



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :

G06F 9/54

A2

(11) International Publication Number:

WO 00/36509

(43) International Publication Date:

22 June 2000 (22.06.00)

(21) International Application Number: PCT/US99/30437

(22) International Filing Date: 17 December 1999 (17.12.99)

(30) Priority Data:

09/21 5,424

18 December 1998 (18.12.98) US

(71) Applicant: UNISYS CORPORATION [US/US]; Township Line and Union Meeting Roads, P.O. Box 500, Blue Bell, PA 19424-0001 (US).

(72) Inventors: GULICK, Robert, C.; 331 Fairview Road, Glenmoore, PA 19343 (US). MORRISSEY, Douglas, E.; 3923 Short Hill Drive, Allentown, PA 18104 (US). CALDARALE, Charles, Raymond; 2843 Brighton Avenue NE, Minneapolis, MN 55418 (US). VESSEY, Bruce, Alan; 916 Noble Drive, Downingtown, PA 19335 (US). RUSS, Craig, F.; 30 Robins Lane, Berwyn, PA 19312 (US). TROXELL, Eugene, W.; 273 Candlebrook Road, King of Prussia, PA 19406-1808 (US). MIKKELSEN, Hans, Christian; 14661 Afton Blvd. South, Afton, MN 55001 (US). MAUER, Sharon, M.; 1064 Westwood Drive, West Chester, PA 19382 (US). CONNELL, Maureen, P.; 2124 Hemlock Road, Norristown, PA 19403 (US). HUNTER, James, R.; 461 East Radnor Ct., Downingtown, PA 19335-1734 (US).

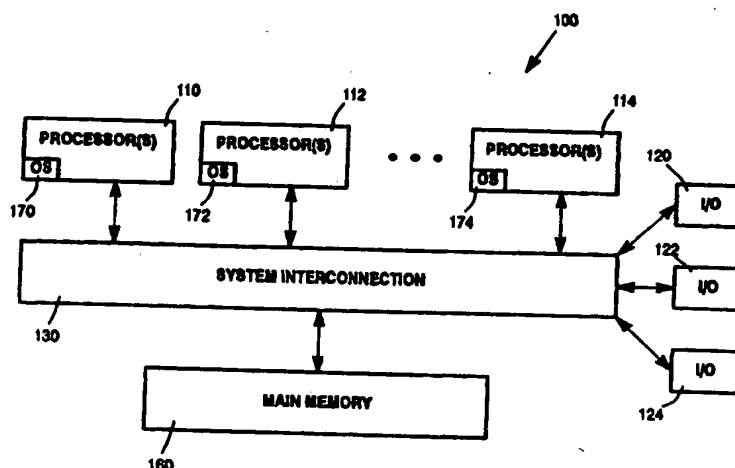
(74) Agents: STARR, Mark, T. et al.; Unisys Corporation, Township Line and Union Meeting Roads, P.O. Box 500, Blue Bell, PA 19424-0001 (US).

(81) Designated States: BR, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

## Published

Without international search report and to be republished upon receipt of that report.

(54) Title: COMPUTER SYSTEM AND METHOD FOR OPERATING MULTIPLE OPERATING SYSTEMS IN DIFFERENT PARTITIONS OF THE COMPUTER SYSTEM AND FOR ALLOWING THE DIFFERENT PARTITIONS TO COMMUNICATE WITH ONE ANOTHER THROUGH SHARED MEMORY



## (57) Abstract

A computer system comprises a plurality of processing modules that can be configured into different partitions within the computer system, and a main memory. Each partition operates under the control of a separate operating system. At least one shared memory window is defined within the main memory to which multiple partitions have shared access, and each partition may also be assigned an exclusive memory window. Program code executing on different partitions enables those partitions to communicate with each other through the shared exclusive memory windows assigned to each partition, so that the exclusive memory windows assigned to each partition appear to the respective operating systems executing on those partitions as if they all start at the same base address.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

**COMPUTER SYSTEM AND METHOD FOR OPERATING  
MULTIPLE OPERATING SYSTEMS IN DIFFERENT  
PARTITIONS OF THE COMPUTER SYSTEM AND FOR  
ALLOWING THE DIFFERENT PARTITIONS TO  
5 COMMUNICATE WITH ONE ANOTHER THROUGH SHARED  
MEMORY**

***Copyright and Trademark Notices***

10 A portion of the disclosure of this patent document contains material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure as it appears in the United States Patent & Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

15 Unix is a registered trademark of The Open Group. SCO and Unixware are registered trademarks of The Santa Cruz Operation, Inc. Microsoft, Window, Window NT and/or other Microsoft products referenced herein are either trademarks or registered trademarks of Microsoft Corporation. Intel, Pentium, Pentium II Xeon, Merced and/or other Intel products referenced herein are either trademarks or registered  
20 trademarks of Intel Corporation.

***Background of the Invention***

25 ***Field of the Invention***

The present invention relates generally to computer systems and, more particularly, to a computer system that operates multiple operating systems in different partitions on the computer system and that allows the different partitions to communicate with one another through shared memory.

30

***Related Art***

A computer system typically includes a processor, main memory, and I/O devices (e.g., printers, network interfaces, graphic display interfaces). The computer system uses an

- 2 -

addressing scheme to designate the source or destination of an item of data. Memory management functions, including the accessing of data, as well as other management functions, are controlled via an operating system. There are a variety of operating systems on the market, each having their own unique characteristics and abilities. Conventional  
5 computer systems typically employ a single operating system.

As modern computer systems grow, and the demands of the user increases, the necessity of employing a plurality of operating systems increases. Unfortunately, a plurality of operating systems substantially increases the complexity of operating the  
10 computer system.

What is needed is a computer system and method for allowing multiple operating systems, including different operating systems, to operate in different partitions on the computer system, and for allowing the different partitions, including the operating systems  
15 and other clients running in the different partitions, to communicate with one another through a shared memory.

### *Summary of the Invention*

20 The present invention is directed to a computer system and methods for allowing multiple operating systems to operate in different partitions within a single computer architecture and for allowing the different partitions to communicate with one another through shared memory.

25 According to a first aspect of the present invention, the computer system comprises a plurality of processing modules and a main memory to which each processing module is connected such that processor-to-memory latency is the same for each processing module across all of the main memory. Groups of one or more processing modules are configured as separate partitions within the computer system,  
30 and each partition operates under the control of a separate operating system. Further

- 3 -

according to this first aspect of the present invention, the main memory has defined therein at least one shared memory window to which at least two different partitions have shared access. Program code executing on different partitions enables those different partitions to communicate with each other through the shared memory  
5 window.

For each different partition configured within the computer system, the main memory may further have defined therein an exclusive memory window to which only that partition has access and in which the operating system of that partition executes.  
10 The separate operating systems on two different partitions may be different operating systems, or may be different instances of a same operating system.

In one embodiment, the program code that enables inter-partition communication (by managing the shared memory window resources) implements a  
15 process by which a sending partition generates an inter-processor interrupt on a receiving partition to signal the receiving partition that information is being transferred to it through the shared memory window. According to this embodiment, the shared memory window comprises a set of input queues associated with each partition, each input queue of the set associated with a given partition corresponding to another  
20 partition and storing entries representing communications from that other partition. In order for one partition (a sending partition) to communicate with another partition (a receiving partition), the program code on the sending partition (i) causes an entry to be created in the input queue of the receiving partition that corresponds to the sending partition; and then (ii) causes an inter-processor interrupt to be generated on the  
25 receiving partition to signal the receiving partition that the entry has been created in that input queue.

Assuming an embodiment in which each partition is assigned only a single interrupt vector for receipt of shared memory inter-processor interrupts from other  
30 partitions, when the inter-processor interrupt is detected on the receiving partition, the program code on the receiving partition (i) causes each of its input queues to be

- 4 -

examined to determine which of the input queues contain entries representing communications from other partitions; and (ii) causes any such entries to be extracted from the input queues that contain them. Preferably, each input queue contains a count of the number of entries in the queue.

5

Alternatively, in an embodiment in which each partition assigns a separate interrupt vector for each other partition from which it may receive an inter-processor interrupt, and wherein the sending partition specifies the interrupt vector assigned to it when sending an inter-processor interrupt to the receiving partition, the receiving  
10 partition can use the specified interrupt vector to identify the input queue associated with the sending partition and process it directly, without having to cycle through all of its input queues (as is the case where each partition assigns only a single interrupt vector for shared memory inter-processor interrupts).

15

Further in accordance with this first embodiment, the shared memory window further comprises a plurality of pages of memory that can be allocated to the partitions, as needed, to facilitate communication of information between them. An input queue entry representing a communication between a sending partition and a receiving partition may comprise a handle to one or more allocated pages of the shared memory  
20 window. A sending partition can use one or more allocated pages to store data representing a message to be communicated to a receiving partition.

25

Still further according to this first embodiment, each input queue is capable of storing a pre-defined number of entries and contains an overflow flag that is caused to  
be set whenever the input queue is full. A sending partition causes the overflow flag of an input queue to be set if the creation of an entry in that input queue causes the input queue to become full. On the receiving side, if a receiving partition encounters an input queue in which the overflow flag is set, it empties the queue and then resets the overflow flag. The receiving partition may then send a communication back to the  
30 sending partition to alert the sending partition that the input queue is no longer full. If an attempt is made to send a communication via an input queue that is full, the sending

- 5 -

partition can return an error, or alternatively, each partition can maintain a location in its exclusive memory window for storing input queue entries that could not be placed in a designated input queue because the overflow flag of that input queue was set previously. The entries stored in the exclusive memory window location can remain  
5 there until the overflow flag of the designated input queue is reset by the receiving partition.

Yet further according to the preferred embodiment, the shared memory window further comprises a table indicating, for each allocable page of the shared memory  
10 window, whether the page is in-use or is available for allocation. The pages that are available for allocation are preferably linked together to form a linked-list of available pages. Ownership of a page by one or more partitions, is preferably indicated, for at least some types of pages by information contained in a header within the page itself. Ownership of other types of pages can be indicated by information in the table that also  
15 specifies the availability of each page.

The header of each page may further comprise a lock field by which one partition may acquire exclusive access to a page in order to, for example, update ownership information in the header of the page. This field is part of a broader lock  
20 mechanism of the present invention that allows different partitions to lock access to the various structures, pages, and tables of the shared memory window, as needed, and in a consistent manner, to ensure that only one partition is capable of modifying any given structure, page, or table at a time (*i.e.*, to synchronize access to these structures). In accordance with one important feature of the lock mechanism of the present invention,  
25 when a memory page is first allocated, the allocating partition must acquire a system wide lock in order to lock access to the page during allocation. However, when ownership of one or more allocated pages is extended or transferred to other partitions, only a lock to the pages involved must be acquired. The lock field in these pages is used for this purpose. This facilitates greater throughput of communications between  
30 partitions, since contention for the system wide lock is eliminated.

- 6 -

According to a second embodiment, the program code on each partition implements a polling process by which each partition polls an area within the shared memory window to determine whether any communications intended for it have been placed in the shared memory window by another partition. In this embodiment, the area that is polled by each partition comprises a plurality of output queues, one for each partition. The output queue for a given partition indicates whether that partition has placed in the shared memory window any communications intended for any of the other partitions. Each partition polls the output queues of the other partitions to determine whether those other partitions have placed any communications intended for it in the shared memory window. Each partition is allocated a separate pool of message buffers in which it may place communications intended for other partitions. When a sending partition places a communication intended for a receiving partition in one of its allocated buffers, it then specifies the location of that buffer in its output queue.

15

In greater detail, the output queue of a given partition comprises one or more node-to-node queues, one associated with each other partition to which it may pass communications. Each node-to-node queue indicates whether communications intended for the partition with which it is associated have been placed in the shared memory window. Thus, each partition polls the node-to-node queues associated with it in the output queues of each other partition to determine whether any of those other partitions have placed any communications intended for it in the shared memory window. For message data that has been placed in a buffer by a sending partition, the node-to-node queue associated with the receiving partition will specify the location of the buffer so that the receiving partition can retrieve the message data.

25

According to a second aspect of the present invention, the computer system may also comprise means for mapping the physical address space of the processors in each partition to the respective exclusive memory window assigned to the partition. Specifically, the means for mapping comprises means for relocating a reference to a location within the physical address space of the processors on a given partition to the

30



- 7 -

corresponding location within the exclusive memory window assigned to that partition. In this manner, the exclusive memory windows of each partition, which are physically located in different areas of the main memory, can be made to appear to their respective operating systems as having a same base physical address in the main memory (*e.g.*,  
5 base address zero). This is necessary in order to run certain off-the-shelf operating systems (*e.g.*, Unix, Windows NT, etc.) in different partitions, because these operating systems assume that main memory starts at address zero. By mapping the processor address space in each partition to its exclusive memory window, the operating systems can continue to reference memory as they normally would in the physical address space  
10 of the processors on which they are executing. Thus, no modification of the operating systems is required.

In a preferred embodiment, the means for relocating comprises a register that holds an offset ( $R_L^{OS}$ ) from the base physical address of main memory to the start of the  
15 exclusive memory window assigned to a given partition, and an adder for adding the offset ( $R_L^{OS}$ ) to each reference by a processor in that partition to a location within its physical address space. As a result, those reference are relocated to their corresponding locations within the exclusive memory window of the partition.

20 According to another feature of this aspect of the present invention, in cases where the physical address space of the processors of a given partition contains a range of addresses unavailable for memory storage (*e.g.*, a range dedicated to memory-mapped I/O), thus defining a memory hole, the computer system may further comprise means for reclaiming for other uses that portion of the exclusive memory window of  
25 the partition that would otherwise correspond to the memory hole. More specifically, the computer system recognizes the memory hole and defines addresses above the memory hole as a high memory range and addresses below the memory hole as a low memory range. In addition to the offset ( $R_L^{OS}$ ) from the base physical address of main memory to the start of the exclusive memory window assigned to a given partition, a  
30 value ( $R_C^{OS}$ ) is also stored that specifies the size of the memory hole. Relocation and reclamation are then achieved by (i) adding the offset ( $R_L^{OS}$ ) to each reference by a

- 8 -

processor in the given partition to a location within the low memory range of its physical address space (thereby relocating those references to their corresponding locations within the exclusive memory window), and (ii) adding the offset minus the value representing the size of the memory hole ( $R_L^{OS} - R_C^{OS}$ ) to each reference by a

5 processor in the given partition to a location within the high memory range of its physical address space (thereby relocating those references to their corresponding locations within the exclusive memory window and at the same time reclaiming that portion of the exclusive memory window that would otherwise have corresponded to the memory hole).

10

According to still another feature of this aspect of the present invention, shared memory windows can also be taken into account. Specifically, as mentioned above, a shared memory window can be defined in addition to the exclusive memory windows for each partition. In order to share access to that window, each partition designates a

15 portion of the physical address space of its processors as corresponding to the shared memory window within the main memory. Then, according to the present invention, the designated portion of the physical address space of the processors on each partition is mapped to the same shared memory window in main memory. In a preferred embodiment, this is achieved in each partition by (i) storing an offset ( $S_{BASE}^{OS}$ ) from the

20 base address of the physical address space of the processors on the partition to the start of the portion of that physical address space designated as corresponding to the shared memory window, (ii) storing another offset ( $S_{BASE}^{MSU}$ ) from the base address of the main memory to the start of the shared memory window within the main memory, and (iii) adding the difference between the offsets ( $S_{BASE}^{MSU} - S_{BASE}^{OS}$ ) to each reference by a

25 processor in the partition to a location within its designated portion, thereby relocating those references to their corresponding locations within the shared memory window of the main memory.

Methods of the present invention are reflected in the various operations of the

30 computer system.

Further features and advantages of the computer system and methods of the present invention, as well as the structure and operation of various embodiments of the present invention, are described in detail below with reference to the accompanying drawings.

## 5 *Brief Description of the Figures*

The invention is best understood by reference to the figures wherein references with like reference numbers indicate identical or functionally similar elements. In addition, the leftmost digits refer to the figure in which the reference first appears in the accompanying drawings in which:

FIG. 1 is a block diagram of an environment suitable for implementation of a preferred embodiment of the present invention;

15           FIG. 2 is a block diagram of a computer system in accordance with a preferred embodiment of the present invention;

FIG. 3 illustrates a view of memory in an example with four partitions, each having an exclusive memory window and access to two shared windows;

FIG. 4 illustrates a view of memory in an example with two partitions each having an exclusive memory window;

FIG. 5 illustrates a view of memory in an example with three partitions, each  
25 having an exclusive memory window and access to one shared window;

FIG. 6 illustrates an example memory configuration that is used to demonstrate the present invention in operation;

FIG. 7 illustrates the result of applying the present invention to the memory configuration shown in FIG. 6;

- 10 -

FIG. 8 is a flowchart illustrating a forward windowing algorithm;

FIG. 9 is a flowchart illustrating a forward translation algorithm;

5

FIG. 10 illustrates an embodiment in which the memory system includes a single shared window, in accordance with the present invention;

FIG. 11 and FIG. 12 illustrate applications of the present invention.

10

FIG. 13 illustrates a process flowchart for an exemplary initialization process, in accordance with the present invention;

FIG. 14 illustrates data structures that can be used for sharing memory, in accordance with a first embodiment of a shared memory management method of the present invention;

15

FIG. 15 illustrates an exemplary embodiment of a message queue area, in accordance with the first embodiment;

20

FIG. 16A illustrates exemplary information that can be included in a node output queue data structure, in accordance with the first embodiment;

FIG. 16B illustrates exemplary information that can be included in a node output queue data structure, in accordance with the first embodiment;

25

FIG. 17 illustrates an exemplary message data structure, in accordance with the first embodiment;

FIG. 18 illustrates an exemplary use of computer system and methods of the present invention for communicating between partitions through shared memory;

30

- 11 -

FIG. 19 illustrates the layout of a shared memory window in accordance with an alternate embodiment of a shared memory management method of the present invention;

5        FIG. 20 illustrates the contents of a control structure header in accordance with the alternate embodiment;

FIG. 21 illustrates the contents of an allocation structure in accordance with the alternate embodiment;

10

FIG. 22 illustrates a block diagram of another exemplary use of the computer system and methods of the present system in which software utilizing the present invention permits operating systems to communicate with shared memory while maintaining an appearance of communication by wire;

15

FIG. 23 illustrates further details of the software illustrated in FIG. 22;

FIG. 24 illustrates further details of the software illustrated in FIG.22, wherein the software is designed to execute in a Windows NT environment;

20

FIG. 25 is a process flowchart illustrating still further details of the software illustrated in FIG.22, wherein the software is designed to execute in a Windows NT environment;

25

FIG. 26 is a process flowchart that illustrates still further details of the software illustrated in FIG.22, wherein the software is designed to execute in a 2200 operating system environment;

30        FIG. 27 is a process flowchart that illustrates still further details of the software illustrated in FIG.22, including details of a co-operative processing communications (CPCOMM) software program;

- 12 -

FIG. 28 illustrates further details of the computer system illustrated in FIG. 2;

FIG. 29 illustrates the contents of an Input Queue Header in accordance with the  
5 alternate embodiment illustrated in FIG. 19;

FIG. 30 illustrates the contents of an Input Queue in accordance with the alternate  
embodiment;

10 FIGS. 31A and 31B comprise a flow diagram further illustrating the operation of  
the computer system in accordance with the alternate embodiment;

FIG. 32A illustrates the contents of a header of a Type 1 shared memory page in  
accordance with the alternate embodiment; and

15

FIG. 32B illustrates the contents of a header of a Type 2 shared memory page in  
accordance with the alternate embodiment.

FIG. 33 is a block diagram of apparatus for carrying out the address relocation and  
20 reclamation methods of the present invention, in accordance with a preferred embodiment  
thereof.

## *Detailed Description of the Preferred Embodiments*

### *Table of Contents*

5	<b>I. Overview</b>
	<b>II. Computer System Platform</b>
	<i>A. Memory Windows (Relocation and Reclamation)</i>
	<i>B. Interleaving and Stacking of Memory (Translation)</i>
	<i>C. Initialization at Boot Time</i>
10	<b>III. Methods for Managing the Global Shared Memory (Inter-Partition Communications)</b>
	<i>A. Polling For Inter-Partition Communications</i>
	<i>B. Interrupt-Driven Shared Memory Communications</i>
15	1. Shared Memory Layout
	2. Free Page List
	3. Client Directory Table
	4. Shared Memory Page Types
	5. Control Structure Header
	6. Allocation Structure
20	7. Signals
	8. Input Queues and Input Queue Header
	9. Inter-Processor Interrupt Mechanism
	10. The Core Services API
	11. Interfaces Supplied by Clients
25	12. Exemplary Operation
	13. Other Functions
	<b>IV. Exemplary Uses of the Computer System and Methods of the Present Invention to Facilitate Communications Between Partitions</b>
	<i>A. A Shared Memory Device Driver</i>
30	<i>B. Maintaining an Appearance of Communications by Wire</i>
	<b>V. Conclusions</b>

- 14 -

## I. Overview

5           The present invention is directed to a multi-processor computer system, having one or more processor modules and a main memory having one or more memory storage units, that allows a plurality of operating systems to concurrently execute in different partitions within the computer system and allows the different partitions to communicate with one another through shared memory. The main memory is divided into a plurality of memory  
10 storage units (referred to as MSU's). The main memory is allocated among the different partitions. Data coherency and consistency are maintained among the partitions.

          According to one inventive aspect of the computer system, an address mapping function,  $f_{pa}$ , is provided between an address request generated from one of the processors  
15 of a processor module and a corresponding address within a window of the main memory. The address mapping function,  $f_{pa}$ , can be conceptually thought of as having three distinct parts: windowing, reclamation and translation.

          The main memory has a contiguous address space. According to the present  
20 invention, each partition (and its associated operating system) is assigned an exclusive memory window within the address space of the main memory. A shared memory window may also be defined within the main memory to which multiple partitions may have shared access. The windowing function maps the physical address space of the processors in each partition to the respective exclusive memory windows assigned to those  
25 partitions. In this manner, the exclusive memory windows of each partition are made to appear to their respective operating systems as having a same base physical address (*e.g.*, address zero) in the main memory. The windowing function is needed to run off-the-shelf operating systems in different partitions on the computer systems, because off-the-shelf operating system (*e.g.*, Unix, Windows NT, etc.) typically expect physical memory to  
30 begin at address zero.



- 15 -

Reclamation reclaims main memory which is located behind memory-mapped I/O address space occupied by, for example, Peripheral Component Interface (PCI), Advanced Programmable Interrupt Controller (APIC) and basic system memory mapped and I/O devices (e.g., floppy controller, serial ports, parallel ports, etc.), which would be unusable by the computer system if not relocated. That is, memory addresses that are allocated to IO devices by each operating system are reclaimed so that the operating system appears to have additional memory space within main memory.

Translation maps a memory reference to a specified memory storage unit. System memory addresses can be interleaved or stacked between memory storage units, as dictated by how the computer system is populated with memory storage units.

In an exemplary embodiment, the computer system includes a plurality of processing modules. A processing module may be a Pod or a sub-Pod. A Pod comprises two sub-Pods. In a preferred embodiment, a maximum configuration of the computer system includes four Pods, *i.e.*, eight sub-Pods. According to the present invention, the computer system can be partitioned on both Pod and sub-Pod boundaries. Thus, in the preferred embodiment, wherein a maximum configuration consists of eight sub-Pods, the computer system can be partitioned into a maximum of eight partitions, each defined by a separate sub-Pod. Further according to the present invention, each partition operates under the control of its own operating system. The operating systems executing on different ones of the partitions may be different operating systems, or different instances of a same operating system.

The invention further provides a global shared memory approach to sharing data between partitions on the computer system. In one embodiment, the global shared memory approach provides an exclusive memory window within the main memory for each partition, plus a shared memory window that multiple partitions can access. The partitions, including their operating systems and/or other clients running within the partitions, can communicate with one another through the shared memory window.

- 16 -

Communications between partitions through shared memory can be managed by any of a variety of methods. In one embodiment, inter-partition communications through shared memory are managed in accordance with an interrupt-driven technique. In another  
5 embodiment, a polling technique is used to manage the shared memory communications.

As used herein, the term "computer system" refers to hardware, including electronic and mechanical components, and to software, including application programs and operating systems. Generally, operating systems include instructions and data that a  
10 computer manipulates in order to perform its tasks. The hardware provides the basic computing resources. The software defines the ways in which these resources are used to solve the computing problems of users.

As used herein, the term "operating system" refers to the program code that  
15 controls and coordinates the use of the hardware among the various application programs for various users. The operating system is the first program code loaded into the main memory of a computer system after the computer system is turned on. The central core of the operating system resides in the memory space at all times. As used herein, the term "operating system address" means the physical address space (memory and I/O) of a  
20 processor of a computer system and is the address space of a conventional computer system as viewed from the perspective of an operating system executing on that computer system.

As used herein, the term "computer architecture" refers to the structure and  
25 behavior of a computer, as viewed by a user. It concerns the specifications of the various functional modules, such as processors and memories, and structuring them together into a computer system. The computer architecture is implemented utilizing hardware.

As used herein, the term "memory storage unit" refers to a memory space capable  
30 of storing information. Each memory storage unit includes a plurality of memory storage units, sometimes referred to as banks of DRAM (Dynamic Random Access Memory). As

- 17 -

used herein, the term "memory storage unit address" refers to an address location as viewed from the perspective of the computer system.

As used herein, the term "partition" refers to one or more processing module(s) that are under the control of a single instance of an operating system. The term "partition" is used herein to refer, in whole or in part, to the processing modules(s) of the partition, the operating system executing on the partition, any exclusive memory window assigned to the partition, other clients or application programs executing on the partition, or any combination thereof.

10

As used herein, the term "processing module" means a plurality of processors operating cooperatively. As exemplified in the preferred embodiment described below, Pods and sub-Pods are both examples of processing modules. One or more Pods or sub-Pods (*i.e.*, one or more processing modules) may be defined as a partition within the computer system.

15

As used herein, the term "program code" means a set of instructions that, when executed by a machine, such as a computer system or processor, causes the computer system or processor to perform some operation. Recognizing, however, that some operations or functionality in a computer system may be hard-coded, in the form of circuitry that performs the operation or function, or may be performed by a combination of executable instructions and circuitry, the term "program code" also includes such circuitry or combination of executable instructions and circuitry.

20

## 25 II. Computer System Platform

Figure 1 illustrates a multi-processor system that includes processor modules 110, 112, and 114. Processor modules 110, 112 and 114 are of comparable compatibility. However, the present invention further contemplates that heterogeneous processors and/or operating systems will co-exist. Each processor module 110, 112 and 114 is self-contained. The processor modules 110, 112 and 114 can each include a plurality of

30

- 18 -

processors. Two or more of processor modules 110, 112 and 114 share access to main (or global) memory 160 and/or to I/O devices 120, 122, and 124, typically through a system interconnection mechanism, such as system interconnection 130. Processor modules 110, 112, and 114 can communicate with each other through main memory 160 (by messages and status information left in common data areas).

According to the present invention, one or more processor modules may be configured as a separate partition within the computer system, such that multiple partitions may exist within the computer system, each partition operating under the control of a separate operating system. For example, each processor module 110, 112 and 114 of Figure 1 can be defined as a separate partition controlled via a separate operating system 170, 172 and 174. Each operating system 170, 172 and 174 views main memory separately as though each is the only entity accessing main memory 160.

A distinction should be made between multi-processor systems and multi-computer systems. A multi-computer system is a system in which computers are interconnected with each other via communication lines to form a computer network. The computers are autonomous and may or may not communicate with each other. Communication among the computers is either via fixed paths or via some message-switching mechanism. On the other hand, a conventional multi-processor system is controlled by one operating system that provides interaction between processors and all the components of the system cooperate in finding a solution to a problem.

Figure 2 is a detailed illustration of a preferred embodiment of a computer system 200, in accordance with the present invention. Computer system 200 includes a main memory, illustrated here as main memory 160, and a plurality of processing modules 240 connected to the main memory via respective third level cache modules 230 and crossbar interconnects 290. In this embodiment, the processing modules and the main memory are arranged in a symmetrical multiprocessing architecture, *i.e.*, processor-to-memory latency is the same for each processing module across all of the main memory.

- 19 -

In the present embodiment, main memory 160 is a directory-based memory system and is capable of supporting various memory consistency models such as, for example, memory consistency models employed on UNIX/NT systems. Main memory 160 includes a plurality of memory storage units (MSUs) 220, such as memory storage  
5 units 220A, 220B, 220C, and 220D. Preferably, each memory storage unit 220A, 220B, 220C, and 220D includes at least eight gigabytes of memory. Preferably, each memory storage unit 220A, 220B, 220C, and 220D includes sixteen semi-independent banks that share four double-wide data busses and eight unidirectional address busses.

10 The plurality of third level cache modules 230, such as third level cache modules 230A through 230D, include a plurality of third level cache application specific integrated circuits (or TCTs), such as TCTs 270A through 270H. In the present embodiment, pairs of processors (*e.g.*, 240A and 240B) share a common bus (*e.g.*, 280A) with a single TCT (*e.g.*, 270A) within a given TLC (*e.g.*, 230A). Each TCT 270 performs address relocation,  
15 reclamation, and translation for memory addresses issued by the processors to which it is connected, as described more fully below.

Each third level cache module 230A through 230D is connected to a respective plurality of processors (MPs) 240A through 240S. Specifically, in the present  
20 embodiment, each TLC 230 is connected to four processors. Each TLC 230 and its respective four processors define a sub-Pod. Further according to the present embodiment, two sub-Pods are connected via a crossbar interconnect (*e.g.*, crossbar interconnect 290A or 290B) to form a Pod. Thus, in the embodiment illustrated in Figure 2, there are four sub-Pods connected via crossbar interconnects 290A and 290B, respectively, to form two  
25 Pods.

Crossbar interconnects 290 interface processors 240, through third level caches 230, with memory storage units 220. Crossbar interconnects 290 employ a crossbar memory approach, whereby a plurality of cross points are placed at intersections between  
30 the processors 240 and memory storage units 220. Within the cross point is a switch that determines the path from a processor bus 280 to a memory storage unit 220. Each switch

- 20 -

point has control logic to set up the transfer path between a processor 240 and main memory 160. The control logic examines the address that is placed on processor bus 280 to determine whether its particular memory storage unit 220 is being addressed. The control logic also resolves multiple requests for access to the same memory storage unit 220 on a predetermined priority basis. Each crossbar interconnect 290 further comprises a pair of Third-Level-Cache Memory Interface application specific integrated circuits (TCMs) 285, which perform address relocation, reclamation, and translation for memory requests from I/O devices, as described more fully below.

Computer system 200 further includes I/O buses 210A through 210D and a plurality of peripheral component interconnects (PCIs), such as PCIs 260A through 260D that are connected via direct I/O bridges, such as direct I/O bridges (DIB) 250A through 250D.

In operation, memory storage units 220 bi-directionally communicate with third level cache modules 230, through crossbar interconnects 290. Crossbar interconnects 290 bi-directionally communicate with direct I/O bridges 250 via I/O buses 210, and with processors 240 through TCTs 270. Direct I/O bridges 250 bi-directionally communicate with peripheral component interconnects 260.

In the present embodiment, the processors (MPs) 240 may comprise Intel processors (*e.g.*, Pentium Pro, Pentium II Xeon, Merced), Unisys E-mode style processors (used in Unisys A Series and Clearpath HMP NX enterprise servers), or Unisys 2200 style processors (used in Unisys 2200 and Clearpath HMP IX enterprise servers. Preferably, a given sub-Pod employs four processors of the same type. However, the present invention contemplates that different sub-Pods may employ different types of processors. For example, one sub-Pod may employ four Intel processors, while another sub-Pod may employ four Unisys E-mode style processors. In such a configuration, the sub-Pod that employs Intel processors may be defined as one partition and may run under the control of an Intel-compatible operating system, such as a version of Unix or Windows NT, while the sub-Pod that employs Unisys E-mode style processors may be defined as another partition

- 21 -

and may run under the control of the Unisys MCP operating system. As yet another alternative, the sub-Pods in two different partitions may both employ Intel processors, but one partition may run under the control of an Intel compatible operating system (*e.g.*, Windows NT), while the other partition may run under the control of the Unisys MCP  
5 operating system through emulation of the Unisys A Series computer architecture on the Intel processors in that partition.

Additional details of the architecture of the preferred embodiment of the computer system 200 of Figure 2 are provided in the foregoing co-pending, commonly assigned  
10 applications listed in the section entitled Cross-Reference to Other Applications, each of which is incorporated by reference herein in its entirety.

As mentioned above, in accordance with the present invention, computer system 200 is partitionable on Pod and sub-Pod boundaries. In figure 28, a portion 2801 of  
15 computer system 200 is illustrated including Pod and sub-Pod boundaries. A Pod 2802 includes crossbar interconnect 290A, a first sub-Pod 2804A, and a second sub-Pod 2804B. Sub-Pods 2804A and 2804B are substantially similar to one another. Sub-Pod 2804A, for example, includes third level cache 230A, which includes TCTs 270A and 270B. Sub-Pod 2804 further includes processors 240A-240D. Pod 2802 thus includes two TLCs 230, four  
20 TCTs 270, eight processors 240 and a crossbar interconnect 290.

In the present embodiment, a maximum configuration of the computer system 200 includes four Pods 2802, each Pod 2802 including two sub-Pods 2804, as described above. Thus, in the maximum configuration, computer system 200 includes (4 Pods) \* (8  
25 processors per Pod) = 32 processors. Computer system 200 can be partitioned on any combination of Pod or sub-Pod boundaries. It is understood, however, that the present invention contemplates other multiprocessing environments and configurations. For example, computer system 200 could be expanded by plugging in more memory storage units 220 and more Pods or sub-Pods.

30

- 22 -

In an embodiment, Pod 2802 is defined to include direct I/O bridges 250A and 250B. In an embodiment, sub-Pods 2804 and 2806 are defined to include direct I/O bridges 250A and 250B, respectively.

5 Further according to the present invention, multiple partitions within the computer system, each of which may comprise one or more Pods or sub-Pods, each operates under the control of a separate operating system. The operating systems executing on the different partitions may be the same or different. For example, the present invention contemplates an environment wherein at least two of the operating systems are different  
10 and one operating system does not control or manage the second operating system.

Figure 5 illustrates an exemplary memory configuration that can be generated on the computer system of Figure 2, in accordance with the partitionability feature of the present invention. In this example, each of three operating systems (OS) has its own  
15 address space 502 (*i.e.*, the physical address spaces of the respective processing modules on which those operating system execute). The main memory 160 has an address space 504. According to the present invention, three exclusive memory windows 540A, 540B and 540C, one for each operating system (*i.e.*, partition), and one shared memory window 537, which is accessible by all three operating systems 540A, 540B and 540C (*i.e.*,  
20 partitions), are defined within the address space 504 of the main memory 160.

For example, OS#1 includes within its address space a low memory window, such as low memory window 511, a low memory hole, such as low memory hole 512, a high memory window, such as high memory window 513, a portion defined as a shared memory  
25 window, such as shared memory window 514, and a high memory hole, such as high memory hole 515. Low memory window 511, low memory hole 512, high memory window 513, and high memory hole 515 are exclusive to operating system OS#1. The portion of the address space defined as the shared memory window 514 is intended to be shared.

30



- 23 -

As used herein, a "high memory hole" refers to memory space in a memory storage unit high address range that is unavailable for storage of data or instructions because the associated address has been assigned to an I/O device. As used herein, a "low memory hole" refers to memory space in a memory storage unit low address range that is  
5 unavailable for storage of data or instructions because the associated address has been assigned to an I/O device. As used herein, a "window" is an address range that has an upper limit and a lower limit. Visibility of and access to a window is governed by ownership rights. As used herein, a "shared window" refers to an address range that at least two operating systems own jointly. That is, more than one operating system has visibility  
10 and access to a shared window. As used herein, the term "exclusive window" refers to an address range which only one operating system owns. That is, only one operating system may view or access an exclusive window. Data coherency and consistency is maintained among operating systems nonetheless.

15 The address space of OS#2 and OS#3 have a similar structure as operating system OS#1. For the sake of brevity, these address spaces will not be described in detail.

The address space of many processors consists of both main memory and memory-mapped Input/Output (I/O) addresses. Main memory transactions are directed to the main  
20 storage units. I/O transactions are forwarded to the I/O subsystem. Since the I/O addresses access additional memory outside of the main storage, the system could end up with a processor address that references two memory locations. For consistency, one of these memory locations will have to be disabled. Disabling these main storage locations creates a hole in the main memory addressing, and results in memory being left unused. If the I/O  
25 memory address space is large, then a significant block of memory is left unusable. If multiple OS partitions are added to the system, then multiple I/O holes are created, resulting in potentially numerous holes scattered across the main memory address space. According to the present invention, as illustrated in Figure 5, low memory holes, such as low memory holes 511, 541, and 571, and high memory holes such as high memory holes  
30 515, 545, and 575, are reclaimed and re-mapped to a contiguous address space, such as is

- 24 -

depicted for MSU memory space 504. MSU memory space 504 is a conceptual view of main memory 160. Reclamation is described below in greater detail.

For example, the contiguous address space of MSU address space 504 includes  
5 low memory, such as low memory 531, 533, and 535, high memory, such as high memory  
532, 534, and 536, and shared memory, such as shared memory 537. Low memory 531  
and high memory 532 comprise an exclusive window exclusive to operating system OS#1.  
Low memory 533 and high memory 534 comprise an exclusive window exclusive to  
operating system OS#2. Low memory 535 and high memory 536 comprise an exclusive  
10 window exclusive to operating system OS#3. There are no memory addressing holes  
within main memory 160. The contiguous address space of main memory 160 is  
maintained independent of memory expansion, type of reference translation (described in  
detail below), or shared memory environment.

#### 15 **A. Memory Windows (Relocation and Reclamation)**

A window is an address range bounded by upper and lower (address) limits.  
Access to and visibility of this space is limited by ownership rights. The present invention  
provides two types of windows: exclusive and shared.

20

Exclusive windows are owned by a single partition/operating system. Every  
instance of an operating system must operate within the limits of its own window. The  
address space of this window is not visible, nor accessible to other partitions/operating  
systems. In a preferred embodiment, all windows begin on a mod 32MB address  
25 boundary. However, other boundaries are contemplated by the present invention. From an  
operating systems point of view, particularly off-the-shelf operating systems such as Unix  
and Windows NT, its address space (*i.e.*, the physical address space of the processor(s) on  
which it executes) always begins at address zero (*i.e.*, its lower limit is zero), as illustrated  
in the left hand portion of Figure 5. From the perspective of main memory 160, the  
30 address range begins at a relocation ( $R_L$ ) value. The  $R_L$  value is described in detail below.

- 25 -

In a preferred embodiment, the upper limit of an exclusive window is set to a base address of a shared window,  $S_{Base}^{OS}$ .

A shared window is an address range bounded by upper and lower limits, where  
5 this space is visible and accessible by more than one operating system (*i.e.*, partition), while each is running within its own exclusive window. The shared window is a common area through which different partitions, including, for example, their operating systems, can communicate and share data. This area also begins on a mod 32MB address boundary in a preferred embodiment. The shared window can be Nx32MB in size. There are two  
10 configuration parameters associated with a shared window. One contains the base address for the portion defined as the shared window within the operating system's address space,  $S_{Base}^{OS}$  (*i.e.*, the base addresses of the portions 514, 544, and 574 for OS#1, OS#2, and OS#3, respectively). The other holds the base address for the corresponding shared area,  $S_{Base}^{MSU}$ , within the address space 504 of main memory 160. In a preferred embodiment,  
15 the upper limit for each operating system's shared area is the "top of memory" value for that operating system. The lower limit,  $S_{Base}^{OS}$ , must be on a mod 32MB address boundary. If exclusive areas are enabled, the location of shared memory 537 within MSU memory space 504 should be above the respective exclusive windows of all the operating systems that share this area. This last requirement is enforced as a hardware design  
20 tradeoff. The shared area is bounded by an upper limit,  $T^{OS}$ , which is an operating system's top of memory reference from within the operating system's addressing viewpoint. An address above  $T^{OS}$  is trapped and never passed to main memory 160. Thus, shared memory 537 is completely bounded.

25 In other configurations contemplated herein, each operating system can coexist with the other operating systems in a totally shared space. An example of this is when an entire MSU block is set to shared. In this case, each operating system can be configured to be able to view the other operating system's address space. When configured in this fashion, the burden of maintaining access rights to individual pages of memory is placed  
30 upon the cooperating operating systems. The hardware no longer restricts accesses and visibility to individual operating systems. The operating systems must control memory

- 26 -

page access rights by processor page control or some other means in order to prevent a process from corrupting memory. This method of operating is used by cooperative operating systems. An operating system can directly read from another operating system's memory page. Also, one operating system instance can load data destined for another operating system directly into the other operating system's data area, by-passing any temporary buffering. Figure 10 illustrates an example of this type of configuration. Referring to Figure 10, each operating system is configured in such a fashion that their shared area provides a view of the entire MSU memory, including a copy of its' own operating system instance. This aliased address is referred to henceforth as a shadow address. The address range residing below the shared area within each operating system's view is referred to as a local address.

In the present embodiment, the present invention limits the association of an exclusive window with a maximum of one shared window. However, in other embodiments, an exclusive window could be associated with more than one shared window. In such a case, there would be separate  $S_{BASE}^{MSU}$  and  $S_{BASE}^{OS}$  values for each such shared window.

According to the present invention, the physical address space of the processing module(s) of each partition (*i.e.*, the address space as viewed by the operating system on that partition) is mapped, or relocated, to the corresponding exclusive memory window assigned to that partition within the address space 504 of the main memory 160. The address space of main memory 160 should be viewed as a single memory block for purposes of discussion. However, the present invention further contemplates a translation function (described below) in which addresses are additionally mapped to an individual memory storage unit 220 in order to produce address interleaving across memory storage units 220.

By way of further example, Figure 4 illustrates a simple system containing two operating systems OS0 and OS1, each occupying 2GB of memory space within main memory 160. Each operating system address space has its own memory-mapped I/O space

- 27 -

415 and 435. In this example, the holes associated with the memory-mapped I/O do not overlay the DRAM memory area.

At this point, the terms Relocation ( $R_L$ ) and Reclamation  $R_C$  can be further  
5 described. Relocation is the assignment of a base address to an exclusive memory  
window. This base address is the starting address (*i.e.*, offset from address zero) for this  
window in the address space of main memory 160 and must be on a mod 32MB address  
boundary. Referring to Figure 4, the  $R_L$  value for operating system window 430 (OS0) is  
zero since the window starts at the bottom of main memory 160. Operating System  
10 window 410 (OS1) has a  $R_L$  value of 2GB, because its physical address zero location has  
been relocated into the address space of main memory 160 starting at 2GB.

Reclamation is the re-mapping of the address space within a window in order to  
reclaim the memory locations that fall behind a memory-mapped I/O address space. If  
15 reclamation is not active and a window has memory-mapped I/O assigned where the I/O  
range falls below the top of memory, a hole is generated in the windows memory address  
space. In the example of Figure 4, reclamation is not needed, because the holes associated  
with the memory-mapped I/O do not overlay the DRAM memory area. However, referring  
to Figure 5, reclamation can be performed for low memory holes 512, 542 and 572 (*i.e.*,  
20 where the 32 bit memory-mapped I/O devices are mapped). Reclamation can be viewed as  
increasing the available memory address space above the hole equal to the size of the hole.  
In a preferred embodiment, reclamation is only performed if the hole size is 128MB or  
larger. This is a hardware tradeoff. Also, because of design tradeoffs, only one address  
hole is reclaimed per operating system instance. The present invention, however,  
25 contemplates that a computer system can be implemented without enforcing these two  
design tradeoffs. Reclamation is discussed in more detail below.

Referring again to Figure 5, all three operating system address spaces OS#1, OS#2  
and OS#3 contain memory-mapped I/O overlaying the memory address space. However,  
30 the low memory hole 512 of operating system address space OS#1 is smaller than the  
minimum 128MB block size, so reclamation is not performed. The low memory hole is

- 28 -

reclaimed for the other two operating systems, however, in their exclusive windows 540A and 540B, respectively.

Figure 3 illustrates another possible configuration containing four operating system windows (or instances). Here OS#1 and OS#4 share a common area, while OS#2 and OS#3 share another. Note that the placement of the individual windows into the address space of main memory 160 is controlled by the  $R_L$  variable. Figure 3 depicts only one of the many possible mappings of these windows into MSU memory space 350.

According to the present embodiment, each operating system window has associated therewith a configuration register that provides a set of configuration parameters:  $R_L^{OS}$ ,  $R_C^{OS}$ ,  $S_{BASE}^{OS}$ , and  $S_{BASE}^{MSU}$ . Different window mappings are easily generated simply by changing the operating system windows' configuration parameters.

TABLE A illustrates the configuration register values for each the operating system windows shown in Figure 5. Reclamation of a memory hole depends on the contents of the configuration register. TABLE A includes a row for each operating system of interest. Relocation field,  $R_L^{OS}$ , stores the base (or starting) address for the operating system window of interest as relocated in the memory storage unit 220. Reclamation field,  $R_C^{OS}$ , stores an address range corresponding to the size of the low memory hole in the operating system window of interest. Shared base OS field,  $S_{BASE}^{OS}$ , stores the base address for the portion of the operating system address space designated as the shared portion. Shared base MSU field,  $S_{BASE}^{MSU}$ , stores the base address for the shared window 537 within the address space of the memory storage unit memory 220.

25

- 29 -

TABLE A

The Configuration Register Values for the Window Mapping Shown in Figure 5.				
	Relocation $R_L^{OS}$	Reclamation $R_C^{OS}$	Shared Base <sup>OS</sup> $S_{BASE}^{OS}$	Shared Base <sup>MSU</sup> $S_{BASE}^{MSU}$
OSOS #1	0.0000.0000 <sub>H</sub> (0GB)	0.0000.0000 <sub>H</sub> (0GB)	1.4000.0000 <sub>H</sub> (5.000GB)	4.5000.0000 <sub>H</sub> (17.250GB)
OSOS #2	1.4000.0000 <sub>H</sub> (5.000GB)	0.1000.0000 <sub>H</sub> (0.250GB)	1.7000.0000 <sub>H</sub> (5.750GB)	4.5000.0000 <sub>H</sub> (17.250GB)
OSOS #3	2.A000.0000 <sub>H</sub> (10.500GB)	0.0800.0000 <sub>H</sub> (0.125GB)	1.B800.0000 <sub>H</sub> (6.87GB)	4.5000.0000 <sub>H</sub> (17.250GB)

In the present embodiment, the TCT 270 for each pair of processors 240 contains the Configuration Register and other registers and logic for performing relocation, reclamation, and translation, as described herein, for addresses issued by the processors interfaced to that TCT. These registers and logic are also replicated in the TCMs 285 of the crossbar interconnects 290, because the TCMs 285 must perform the same relocation, reclamation, and translation on memory requests received from an I/O processor (e.g., PCI card) via a respective DIB 250.

10

Within the physical address space of the processors of each partition, the TCTs 270 of that partition determine an address range for low memory, high memory, low memory holes, high memory holes, and shared memory. For example, in the address space of operating system OS#3, low memory window 571 begins at address location 0.000<sub>H</sub> and includes 3.875 gigabytes of memory space. High memory window 573 begins at address location 1.5000.000<sub>H</sub> and includes 5.250 gigabytes of memory space. Low memory hole 572 includes 125 megabytes of unused memory space to be reclaimed. High memory hole 575 includes 250 megabytes of unused memory to be reclaimed.

15

- 30 -

In performing its windowing function, each TCT 270 of the present invention further assigns its partition an exclusive memory window within the address space 504 of the main memory 160. Within each exclusive memory window, there is an address range for low memory and for high memory. For example, in exclusive window 540B, low memory window 533 begins at address location 1.4000.0000<sub>H</sub> and includes 5.000 gigabytes of memory space. High memory window 534 begins at address location 2.8000.000<sub>H</sub> and includes 10.000 gigabytes for a total of 10.500 gigabytes of memory space in exclusive window 540B. In exclusive window 540A, low memory window 535 begins at address location 2.A000.0000<sub>H</sub> and includes 5.125 gigabytes of memory space. High memory window 534 begins at address location 3.E800.000<sub>H</sub> and includes 1.625 gigabytes of memory space.

When one of the processors of a processing module of a given partition issues an address on its address lines ("the referenced address" or "processor address"), the TCT 270 for that processor adjusts the address for any relocation, reclamation, or shared windowing, as required, to produce the address of the corresponding location in the main memory 160. The values in the various fields of the configuration register (TABLE A) are used during this process. Specifically, if the referenced address is within the portion of the operating system address space designated as the shared window, then the referenced address is offset by the values contained in shared base OS field and shared base MSU fields of the configuration register. If the referenced address is within the high memory window of the operating system's address space, then the referenced address is offset by the values contained in the relocation and reclamation fields of the configuration register. If the referenced address is within the low memory window of the operating system's address space, then the referenced address is offset by the value contained in the relocation field of the configuration register. As described herein, therefore, the TCTs 270 provide a means for mapping the physical address space of the processors in each partition to the respective exclusive memory windows assigned to each partition, and, more specifically, a means for relocating a reference to a location within the physical address space of the processors on a respective partition to the corresponding location within the exclusive memory window assigned to that partition. As mentioned above, in a similar manner, the TCMs 285



- 31 -

perform any relocation or reclamation required for memory addresses received from an I/O processor (e.g., PCI card) communicating via a DIB and TCM to main memory.

TABLE B illustrates pseudo-code for implementing relocation and reclamation of operating system address spaces (*i.e.*, the physical address spaces of the processors of the different partitions) to their corresponding exclusive memory windows within main memory. Generally, memory-mapped I/O addresses are filtered out by the TCT 270, leaving only references to main memory 160. The remaining addresses are then passed through the algorithm shown in TABLE B, as described in detail below. Finally, the relocated memory reference is passed to main memory 160.

TABLE B

```

if  $OS_{ADR}^{OS} \in RANGE_{SHAREDMEMORY}$ 
then  $MSU_{ADR} = OS_{ADR}^{OS} + [S_{BASE}^{MSU} - S_{BASE}^{OS}]$ ;
elseif  $OS_{ADR}^{OS} \in RANGE_{HIGHMEMORY}$ 
then  $MSU_{ADR} = OS_{ADR}^{OS} + [R_L^{OS} - R_C^{OS}]$ ;
else /*  $OS_{ADR}^{OS} \in RANGE_{LOWMEMORY}$  */
 $MSU_{ADR} = OS_{ADR}^{OS} + [R_L^{OS}]$ ;
endif;
```

Figure 8 illustrates a flow chart of the address windowing algorithm. Reference is also made to TABLE A. As shown in step 810, a check is performed to determine whether a reference address (*i.e.*, an address issued by one of the processors of a processing module within a given partition executing a given operating system),  $OS_{ADR}$ , is within the portion of the operating system's address space designated as the shared memory window. If so, the referenced address is relocated to an address based on the formula:  $OS_{ADR} + [S_{BASE}^{MSU} - S_{BASE}^{OS}]$ , as shown in step 815. This is referred to as the relocated address, which in turn is used to access main memory 160. The relocated address is the address of the

- 32 -

corresponding location in the shared memory window defined within the main memory 160.

Otherwise, a check is performed to determine whether the referenced address is within the high memory portion of the operating system address space (e.g., high memory 513, 543 or 573). This is shown in step 820. If so, the referenced address is relocated to an address based on the formula:  $OS_{ADR} + [R_L^{OS} - R_C^{OS}]$ , as shown in step 825. The relocated address identifies the corresponding location in the exclusive memory window for the partition.

10

Otherwise, the algorithm assumes that the referenced address falls within the low memory portion of the operating system address space (e.g., low memory 511, 541 or 571), as shown in step 830. In this case, the referenced address is relocated to an address based on the formula:  $OS_{ADR} + [R_L^{OS}]$ . Thus, address references within the physical address space of a processor within a partition (*i.e.*, the address space viewed by the operating system) are relocated to their corresponding locations within either the exclusive memory window defined for that partition within main memory or the shared memory window defined within main memory.

15

20

Figure 33 is a block diagram illustrating apparatus, in the form of registers and logic, for performing the relocation and reclamation functions described above, in accordance with the preferred embodiment. This logic is provided in each TCT 270 to perform the relocation and reclamation functions of the present invention for memory addresses issued by the processors (MP) 240 interfaced to the TCT 270. As mentioned, this logic is also replicated in each TCM 285 in order to perform relocation and reclamation for memory addresses issued by an I/O processor via a respective DIB 250.

25

According to the preferred embodiment, as illustrated in Figure 33, a memory address issued on the address lines of a given processor 240 (or by an I/O processor via a respective DIB 250) is captured in a Processor\_Address register 3310. In the preferred embodiment, main memory is addressable in words of 8 bytes bits (1 word = 8 bytes = 64

30

- 33 -

bits), and therefore, the least significant 3 bits of the processor address are not needed for generating an adjusted address. Thus, as shown, only bits [35:3] are captured in the Processor\_Address register 3310. Furthermore, in the preferred embodiment, main memory is cached in blocks of eight (8) words (8 words = 64 bytes), and thus bits [35:6] represent the effective cache block address. As shown, these bits are captured in a subsequent Cache\_Block\_Address register 3312.

As further described above, in the preferred embodiment, all memory windows, whether "exclusive" or "shared," must begin on a mod 32MB address boundary. Consequently, in relocating a processor address to a particular exclusive memory window or shared memory window, only bits [35:25] of the processor address are needed for the calculation. Accordingly, as shown, these bits are captured to a temporary register 3314.

The values  $S_{BASE}^{MSU}$ ,  $S_{BASE}^{OS}$ ,  $R_L^{OS}$ , and  $R_C^{OS}$  are stored in respective register locations 3318, 3320, 3330, and 3340. Collectively, these register locations comprise the Configuration Register described above. In practice, these register locations can comprise separate fields of a single, larger register, or they can be implemented as four separate registers. For the case of a processor address that falls within the portion of the processor's address space designated as a shared memory window, a subtractor 3405 subtracts the  $S_{BASE}^{OS}$  value in register location 3320 from the  $S_{BASE}^{MSU}$  value in register location 3318 and stores the resulting offset value in register 3350. For the case of a processor address that falls within the high memory portion of the exclusive memory window assigned to the partition to which the processor belongs, a subtractor 3410 subtracts the  $R_C^{OS}$  value in register 3340 from the  $R_L^{OS}$  value in register 3330 and stores the resulting offset value in register 3370. As further shown, the five bits of the  $R_C^{OS}$  value are padded (using an append function 3400) with two logic zero bits in the least significant bit positions and four logic zero bits in the most significant bit positions to properly align the bits for subtraction from the bits of the  $R_L^{OS}$  value. Recall from above that in the present embodiment, reclamation can only be performed in increments of 128 MB. For the case of a processor address that falls within the low memory portion of the processor's exclusive memory

- 34 -

window, the  $R_{i,OS}$  value in register 3330 is the required offset, and thus, that value is stored directly in register 3360.

Address Range Compare Logic 3390 performs the steps described above of  
5 determining whether the address issued by the processor falls within the portion of the  
processor's address space designated as a shared memory window, or whether the address  
falls within either the low memory or high memory portions of the exclusive memory  
window assigned to the partition to which the processor belongs. Based on this  
comparison, the appropriate offset from one of the registers 3350, 3360, or 3370 is selected  
10 via a 3:1 Selector 3380. An adder 3420 then adds the selected offset value to the bits  
[35:25] of the processor address stored in register 3314, and the result is stored in register  
3430. The bits in register 3430 are then prepended to bits [24:6] of the cache block address  
to form the adjusted address, which is stored in an Adjusted\_Partition\_Address register  
3316. The adjusted address in register 3316 is then used to access main memory (after  
15 further translation in accordance with the interleaving mechanism of the present invention  
described below).

Referring again to Figure 5, and as already discussed above, addresses that have  
been assigned to memory-mapped I/O can be reclaimed. These addresses are referred to as  
20 low memory holes, such as low memory hole 512. In a preferred embodiment, the low  
memory holes always begin immediately below 4GB and extend downward in the address  
space of the associated operating system equal to the size of the hole. Obviously the  
placement of the low memory hole is a design choice. Memory reclamation is to be used  
only when the top of memory addresses, for the installed memory amount, is greater than  
25 the bottom of the memory overlap region (i.e., 4GB minus the overlap hole size). In other  
words, reclamation should not be used in systems where there is no overlap between the  
PCI APIC range and installed DRAM memory.

All overlaid memory, and any memory immediately above it, can be perceived as  
30 sliding up in the processor/operating system address space. Therefore, the memory that  
lies behind and starting at the bottom of the hole now begins at address 4GB and extends

- 35 -

upward from this point. Memory addressing remains contiguous from the 4GB starting address and extends to the new top of memory, i.e., the original top of memory plus the hole size.

- 5           Figure 11 illustrates how an address range is mapped using a specific example. For systems with 4GB or less of memory and where there is a partial memory overlay with the PCI APIC Range, reclamation can be used. In these systems, the overlapped memory is mapped to start at 4GB. Figure 12 illustrates this point. The sub-Pod takes a processor's adjusted memory request address, and after determining that it lies above the 4GB
- 10   boundary, it subtracts a fixed value from it. This memory address reflects the insertion of the PCI APIC Range into the system address space. Therefore, the adjustment offset is equal to the PCI APIC Range hole size, fixed in increments of 128MB blocks as described above.
- 15           Provided below are a few more examples of relocation and reclamation in accordance with the present invention. Reference is made to Figure 5 and TABLE A. The first example deals with an address reference within an exclusive window. The second example references a shared window.
- 20           As shown in Figure 5, operating system address space OS#3 has been relocated ( $R_L$ ) to main memory address 10.5 GB. Reclamation is set to recover the 128MB (0.125GB) memory behind the low memory hole 572. Using  $OS_{ADR} = 1.5000.0000_H$  as the memory reference, TCT 270 performs the function  $OS_{ADR} + [R_L - R_C]$  to generate an address in MSU memory space 504. The values for  $R_L$  and  $R_C$  are provided in TABLE A.
- 25   Thus,  $OS_{ADR} + [R_L - R_C]$  becomes  $1.5000.0000_H + [2.A000.0000_H - 0.0800.0000_H]$ . This becomes  $1.5000.0000_H + 2.9800.0000_H$ , which becomes  $3.E800.0000_H$  (15.625 GB). This address corresponds to a location within exclusive window 540A, which is associated with operating system OS#3. A simple calculation shows the address is offset 1.25GB from high memory area base address of 4GB. The address calculated above is also offset
- 30   1.25GB from the relocated high memory base address (14.375GB) of OS #3.

- 36 -

If a processor in the partition in which OS#2 is executing issues the same address, 1.5000.0000<sub>H</sub>, the relocated address will instead fall within the exclusive memory window assigned to that partition (i.e., window 540B). Thus  $OS_{ADR} + [R_L - R_C]$  becomes 1.5000.0000<sub>H</sub> + [1.4000.0000<sub>H</sub> - 0.1000.0000<sub>H</sub>]. This becomes 1.5000.0000<sub>H</sub> + 1.3000.0000<sub>H</sub>, which becomes 2.8000.0000<sub>H</sub> (10.00GB). This address clearly falls in high memory area 534 of main memory 160, which is part of the exclusive memory window (540B) assigned to the partition executing OS#2. This example demonstrates that the operating systems in two different partitions will each view their address spaces as if starting at the same base address (i.e., address zero), but address references within those address spaces will be properly relocated to their corresponding locations within the exclusive memory windows assigned to each partition within main memory. Of course, the relocation feature of the present invention can be used to map any two overlapping physical address spaces on different partitions (not just those that both start at address zero) to the respective exclusive memory windows in main memory.

15

The second example uses memory references to shared window 575 associated with OS#3. For this example, assume OS#3 tries to reference address 1.B900.0000<sub>H</sub> (6.890GB). TCT 270 determines that this address falls within the range of shared memory. As such, the present invention applies the function mapping  $OS_{ADR} + [S_{BASE}^{MSU} - S_{BASE}^{OS}]$  to generate an appropriate address to access MSU memory space 504. Thus the mapping function becomes 1.B9000.0000<sub>H</sub> + [4.5000.0000<sub>H</sub> - 1.B8000.0000<sub>H</sub>]. This becomes 1.B9000.0000<sub>H</sub> + 2.98000.0000<sub>H</sub>, which becomes 4.5100.0000<sub>H</sub> (17.2656GB). This address falls within the range of shared memory window 537 of MSU memory space 504.

25

Using the same address offset, 0.0156GB, and applying it to operating system OS#2's shared base address, the equivalent address can be calculated for OS#2.  $OS_{ADR}$  equals 5.750GB + 0.0156GB, which equals 5.7656GB (1.7100.0000<sub>H</sub>). Applying the mapping function,  $OS_{ADR} + [S_{BASE}^{MSU} - S_{BASE}^{OS}]$ , we get 1.7100.0000<sub>H</sub> + [4.5000.0000<sub>H</sub> - 1.7000.0000<sub>H</sub>]. Thus the mapping function generates a memory address of 4.5100.0000<sub>H</sub> (17.2656GB). Thus, a memory reference by operating system OS#3 of 1.B900.0000<sub>H</sub>

- 37 -

(6.8906GB) and a memory reference by operating system OS#2 of 1.7100.0000<sub>H</sub> (5.7656GB) both access main memory 160 at address 4.5100.0000<sub>H</sub> (17.2656GB).

### ***B. Interleaving and Stacking of Memory (Translation)***

5

Translation is the process by which a memory reference (after relocation and, if appropriate, reclamation) is mapped to a specific memory storage unit within main memory 160. Referring to Figure 2, main memory 160 is conceptually divided into a plurality of MSU pairs 222 and 224 (referred to as MSU\_PAIRs). Individual MSU's 220 within a MSU\_Pair are not uniquely connected. Only two MSU\_PAIRs 222, 224 are shown in Figure 2 for illustration purposes only. The present invention contemplates more than two MSU\_PAIRs.

Computer system 200 utilizes the adjusted address (or memory reference) that was generated during relocation and, if applicable, reclamation as described above, and then interleaves or stacks the adjusted memory reference between memory storage unit pairs 222, 224. The goal of the present invention is to distribute each of the main memory requests associated with each processor 240 over the global address space of main memory 160 (i.e., total DRAM address space) such that sequential memory accesses are distributed over different memory storage units 220 in order to minimize contention for memory resources. In the event interleaving cannot be performed, memory addresses are directed to memory storage unit pairs in a sequential order, referred to herein as stacking.

In an exemplary embodiment, there are four memory storage units, i.e., two pairs of memory storage units, such as memory storage unit pair 222 and memory storage unit pair 224. Each memory storage unit pair (hereinafter MSU\_Pair) includes two memory storage units, such as memory storage units 220A and 220B. Interleaving is accomplished across memory storage unit pair 222 and 224. Then, interleaving is accomplished across the memory storage units 220 within the memory storage unit pairs 222 and 224, respectively. There effective result is four-way interleaving.

- 38 -

For example, suppose there are two memory storage units, such as memory storage unit 220A and memory storage unit 220B. Optimally, references to memory would be ping-ponged between memory storage unit 220A and memory storage unit 220B. That is, the first reference to memory accesses memory storage unit 220A, while the second accesses memory storage unit 220B. If memory storage unit 220A has only one bank populated, while memory storage unit 220B has eight banks populated, ping-ponging between memory storage unit 220A and memory storage unit 220B, then at some point memory storage unit 220A will run out of memory space. In that case, the remaining memory in memory storage unit 220B will be stacked, i.e., resort to sequential addressing (or referencing) of memory storage unit 220B.

One characteristic of memory storage units is that there may be one memory storage unit present or a plurality of memory storage units present in a particular memory storage unit "pair." Moreover, memory storage units can be populated at different rates. That is, one memory storage unit can have one bank of DRAM populated, while another memory storage unit may have eight banks of DRAM populated.

In accordance with the present invention, the translation process involves interleaving and stacking of memory references between memory storage unit pair 222 and memory storage unit pair 224, and among MSUs 220. For a memory request issued from a processor (MP) 240, this process is performed by the respective TCT 270. For memory requests issued from an I/O processor (*e.g.*, PCI card) via a DIB, this process is performed by the respective TCM 285.

25

Considering the operation of a TCT 270, a mechanism is provided for specifying at initialization time which MSU\_Pair or which MSU 220 should receive the first cacheline address (*i.e.*, an address from the TCT 270). The TCT 270 takes a processor's memory read/write address (after any relocation and/or reclamation) and passes it through an address translation function. In a preferred embodiment, memory storage unit 220 receives a twenty-eight bit cache line address (or memory reference) and an 8 byte container

30



- 39 -

address from a multi-cycle signal representing 16 gigabytes of memory space. Based on the settings of the address translation options, which are described below, the translation function generates a MSU number that is associated with the memory storage unit that will receive the request, along with the upper ten 10 bits of the 28 bit MSU mapped address.

- 5 The TCT 270 also provides the MSU's lower 18 bits of the mapped address; however, these bits are not altered by the translation function.

- A TCT 270 allows for various combinations of interleaving and stacking of memory accesses on both a MSU\_Pair basis and between each individual MSU 220.
- 10 Listed in TABLE C are the eight combinations for interleaving/stacking memory between MSU\_PAIRs and their individual MSU's 220.

**TABLE C**

Option	Between MSU_Pair0 & MSU_Pair1	MSU_Pair0 Between MSU0 & MSU1	MSU_Pair1 Between MSU2 & MSU3
ISS	Interleaved	Stacked	Stacked
ISI	Interleaved	Stacked	Interleaved
IIS	Interleaved	Interleaved	Stacked
III	Interleaved	Interleaved	Interleaved
SSS	Stacked	Stacked	Stacked
SSI	Stacked	Stacked	Interleaved
SIS	Stacked	Interleaved	Stacked
SII	Stacked	Interleaved	Interleaved

- 40 -

Referring to TABLE C, in the III mode, the algorithm distributes every other cache line to alternating MSU\_PAIRS (e.g., cache line address 0 forwarded to MSU\_PAIR 222).

The algorithm further distributes every other cache line directed to an MSU\_PAIR to  
5 alternating MSUs 220 in MSU\_PAIR 222, 224 (e.g., cache line address 0 is directed to the lower numbered MSU 220).

In ISI, ISS or IIS mode, the algorithm distributes every other cache line to alternating MSU\_PAIRS 222, 224 (e.g., cache line address 0 is forwarded to MSU\_PAIR  
10 222). For MSUs 220 within a MSU\_PAIR 222, 224 that are stacked in accordance with the present invention, the algorithm further directs sequentially addressed accesses to the lower numbered MSU 220 of the selected MSU\_PAIR 222, 224 until it is full before sequentially filling the other MSU 220. For MSUs 220 within a MSU\_PAIR 222, 224 that are interleaved in accordance with the present invention, the algorithm further distributes  
15 every other cache line directed to a MSU\_PAIR 222, 224 to alternating MSUs 220 (i.e., cache line address 0 is directed to the lower numbered MSU 220 within MSU\_PAIR 222, 224).

In SSS mode, the present invention sequentially fills the lower numbered MSU\_PAIR 222, 224 (determined by a configuration register) until it is full before  
20 sequentially filling the other MSU\_PAIR 222, 224. The algorithm further directs accesses sequentially to the lower numbered MSU 220 within the selected MSU\_PAIR 222, 224 until it is full before sequentially filling the other MSU 220 of that MSU\_PAIR 222, 224.

In SSI, SII or SIS mode, the algorithm sequentially fills the lower numbered  
25 MSU\_PAIR 222, 224 until it is full before sequentially filling the other MSU\_PAIR 222, 224. For MSUs 220 within a MSU\_PAIR 222, 224 that are stacked, the present invention then sequentially addresses the low MSU 220 of the selected MSU\_PAIR 222, 224 until it is full before sequentially filling the other MSU\_PAIR 222, 224. For MSUs 220 within a MSU\_PAIR 222, 224 that are interleaved, the present invention distributes every other  
30 cache line in a MSU\_PAIR 222, 224 to alternating MSUs 220. Cache line address 0 is directed to the lower numbered MSU 220 within that MSU\_PAIR 222, 224.

- 41 -

For example, following the ISS option, interleaving is accomplished every other cache line to alternating memory storage unit pairs. That is, a first cache line address is forwarded to memory storage unit pair 222 and the next cache line address is forwarded to memory storage unit pair 224. The present invention sequentially stacks memory references in memory storage unit 220A until memory storage unit 220A is full. When memory storage unit 220A is full, the present invention then sequentially stacks memory references in memory storage unit 220B until it is full. Similarly, when memory storage unit 220C is full, the present invention then sequentially stacks memory references in memory storage unit 220D until it is full.

TABLE D defines a translation and reclamation register. The table includes a row for each address bit of interest within the translation and reclamation register. Each row includes a function field and a default value field. Function field indicates the function of the address bit of interest. Default value field is the value that the address bit defaults to upon initialization. The status of the bits in memory address translation and reclamation register determine whether memory address space reclamation is enabled and whether address translation is enabled. It also indicates which memory storage unit pair to select and which memory storage unit to select for the translation process.

**TABLE D**

Bits	Function	Default Value
[15]	Address Translation Enable	0 (Default)
[14]	Memory Address Space Reclamation Enable	0
[13]	PAIR_MODE	0
[12]	PAIR_SEL	0
[11:10]	Reserved	00
[9:0]	Smallest_Pair_Size[9:0]	000H (Default)

- 42 -

It is the responsibility of a memory controller (not shown) to interleave between banks of an MSU\_PAIRs 222, 224 and MSUs 220.

- 5 Whether computer system 200 implements interleaving depends on the settings in a plurality of registers. For example, TABLES E and F illustrate the contents upon initialization of a memory address translation register corresponding to a first memory storage unit pair and a second memory storage unit pair, respectively. Memory address translation register includes a row for each bit of interest. Each row includes a function
- 10 field and a default value field. Function field includes the function of the address bit of interest. Default value field is the value that the address bit defaults to upon initialization.

**TABLE E**

Bits	Function	Default Value
[15]	Pair#0 Address Translation Enable	0 (Default)
[14]	Reserved	0
[13]	Pair0_Mode	0
[12]	Pair0_Sel	0
[11:10]	Reserved	00
[9:0]	Pair0_Smallest_MSU_Size[9:0]	000H (Default)

15

**TABLE F**

Bits	Function	Default Value
[15]	Pair#1 Address Translation Enable	0 (Default)
[14]	Reserved	0
[13]	Pair1_Mode	0
[12]	Pair1_Sel	0
[11:10]	Reserved	00
[9:0]	Pair1_Smallest_MSU_Size[9:0]	000H (Default)

The status of the bits in memory address translation registers shown in TABLE E and F determine whether interleaving for a particular pair of memory storage units is enabled or whether stacking is enabled. The status of the bits in memory address translation registers further indicate the smaller of the two memory storage units in a memory storage unit pair.

TABLE G shows Configuration Information required at initialization for forward and reverse address translation. TABLE G relates to Figure 2 as follows: MSU\_Pair0 is MSU\_Pair 222, MSU\_Pair1 is MSU\_Pair 224, MSU#0 is MSU 220A, MSU#1 is MSU 220B, MSU#2 is MSU 220C and MSU#3 is MSU 220D.

**TABLE G**

Name	Definition
<b>MSU_Pair0/Pair1 Configuration Registers: used to control accesses to a specific MSU_Pair</b>	
PAIR_MODE	This 1 bit register controls whether address interleaving between MSU_PAIRs is selected. Address interleaving should only be enabled when both MSU_PAIRs are present. When

- 44 -

Name	Definition
	= 0 then Interleave between MSU_PAIRs = 1 then Stack between MSU_PAIRs (Pair0 first, overflow into Pair1)
SMALLEST_PAIR_SZ	This register <sup>1</sup> holds one of two memory size values <sup>2</sup> depending on whether address interleaving between MSU_PAIRs is enabled. if PAIR_MODE = 0 (interleaving then) = the smaller of the two memory size values between MSU_Pair0 (MSU#0 + MUS#1) and MSU_Pair1 (MSU#2 + MSU#3). else PAIR_MODE = 1 (stacking) = the memory size of the MSU pairs selected by the PAIR_SEL register
PAIR_SEL	This 1 bit register specifies which one of the two MSU_PAIRs is to be addressed first. The value depending on whether address interleaving is being performed. For interleaving, the MSU_Pair with the largest installed memory must be selected. For stacking, either MSU_Pair can be selected. if PAIR_MODE = 0 (interleaving) then = 0 if pair0 has more storage then pair1 = 1 if pair1 has more storage then pair0 else PAIR_MODE = 1 (stacking) = Pair which gets the memory "address0" (0 - Pair0; 1 - Pair1)
<b>MSU_Pair0 Configuration Registers: used to control accesses to a specific MSU within pair0</b>	
PAIR0_MODE	This 1 bid register controls whether address interleaving between MSUs within an MSU_Pair is selected. Address interleaving should only be enabled when both MSUs are present in MSU_Pair0. = 0 Interleave between MSUs of pair0 (MSU#0 and MSU#1) = 1 Stack the MSUs of pair0
PAIR0_SMALLEST_MSU_SZ	This register <sup>1</sup> holds one of two memory size <sup>2</sup> values depending on whether address interleaving within this MSU_Pair is enabled. = the smaller of the two memory size values between MSU#0 and MSU#1 of MSU_Pair0.

- 45 -

Name	Definition
	<p>else (PAIR0_MODE0 = 1:stacking)</p> <p>= the memory size of the MSU selected by the PAIR0_SEL register</p>
PAIR0_SEL	<p>This 1 bit register specifies one of the two MSUs within a MSU_Pair is to be addressed first. The value depending on whether address interleaving is being performed. For interleaving, the MSU with the largest installed memory must be selected. For stacking, either MSU can be selected.</p> <p>if PAIR0_MODE = 0 (interleaving) then</p> <p>= 0 if MSU#0 of pair0 has more storage then MSU#1 of pair0</p> <p>= 1 if MSU#1 of pair0 has more storage then MSU#0 of pair0</p> <p>else PAIR0_MODE = 1 (stacking)</p> <p>= MSU of pair0 which gets the memory "address 0"</p> <p>(0 - MSU#0; 1 - MSU#1)</p>
<b>MSU_Pair1 Configuration Registers: used to control access to a specific MSU within pair1</b>	
PAIR1_MODE	<p>This 1 bit register controls whether address interleaving between MSUs within an MSU_Pair is selected. Address interleaving should only be enabled when both MSUs are present in MSU_Pair1.</p> <p>When</p> <p>= 0 Interleave between MSUs of pair1 (MSU#2 and MSU#3)</p> <p>= 1 then Stack the MSUs of pair1</p>
PAIR1_SMALLEST_MSU_SZ	<p>This register<sup>1</sup> holds one of two memory size values<sup>2</sup> depending on whether address interleaving within this MSU_Pair is enabled.</p> <p>if PAIR1_MODE = 0 (interleaving) then</p> <p>= the smaller of the two memory size values between MSU#2 and MSU#3 of MSU_Pair1.</p> <p>else PAIR1_MODE = 1 (stacking)</p> <p>= the memory size of the MSU selected by the PAIR1_SEL register</p>
PAIR1_SEL	<p>This 1 bit register specifies one of the two MSUs within a MSU_Pair is to be addressed first. The value depending on whether address interleaving is being performed. For interleaving, the MSU</p>

- 46 -

Name	Definition
	<p>with the largest installed memory must be selected. For stacking, either MSU can be selected.</p> <p>if PAIR1_MODE = 0 (interleaving) then</p> <p style="padding-left: 40px;">= 0 if MSU#2 of pair1 has more storage then MSU#3 of pair1</p> <p style="padding-left: 40px;">= 1 if MSU#3 of pair1 has more storage then MSU#2 of pair1</p> <p>else PAIR1_MODE = 1 (stacking)</p> <p style="padding-left: 40px;">= MSU of pair1 which gets the memory "address 0"</p> <p style="text-align: right;">(0 - MSU#2; 1 - MSU#3)</p>

<sup>1</sup>Note: The size of this register is not specified in this table. It is implementation specific, and is not necessary for the understanding of the translation algorithm.

- 5     <sup>2</sup>Note: The memory size is equal to the maximum memory address + 1. For example, a single 128MB bank has an address range from 000\_0000<sub>H</sub> - 700\_0000<sub>H</sub>, but the size is 800\_0000<sub>H</sub>. Expanding this size to 36 bits [35:0] yields 0\_800\_0000<sub>H</sub>. Using the 9 most significant bits [35:27] for the size, the size register for this example is loaded with 000000001<sub>B</sub> or 001<sub>H</sub>.

10         As mentioned, logic and registers to implement the forward address translation function reside in both the TCMs 285 (for memory requests from an I/O processor via a respective DIB) and the TCTs 270 (for memory requests from a processor 240). The algorithm is performed in two steps. The first step determines which MSU\_PAIR should be selected and the second step determines which MSU of the selected pair should be selected to send the address to. Illustrated in Appendix A is simplified pseudo-code of the forward address translation algorithm. The pseudo-code does not include checks verifying criteria such as the number of MSU\_PAIRS, or the number of MSUs per MSU\_PAIR, etc.

15         These checks, which should be readily apparent to one skilled in the art, were intentionally left out of the pseudo-code allowing for easier understanding of the translation process.

20

The forward address translation algorithm takes as an input TEMP\_ADDR and uses registers PAIR\_MODE, SMALLEST\_PAIR\_SZ and PAIR\_SEL. The algorithm produces as an output TEMP\_ADDR, which is the address after any required adjustments, and RCVING\_PAIR, which indicates which MSU\_PAIR has been selected. Initially,



- 47 -

TEMP\_ADDR [29:0] is the address after any address relocation has been performed. TEMP\_ADDR [29:0] equals ADDR\_IN [35:6]. TOP\_OF\_INTRLV\_RANGE is the address value where there is no more memory left for interleaving. That is, this is the address where stacking of memory addresses begins. TOP\_OF\_INTRLV\_RANGE equals  
5 two times SMALLEST\_PAIR\_SZ.

Figure 9 illustrates a flowchart of the forward address translation algorithm. The selection of an MSU\_Pair is shown in stage 900. Step 902 determines whether interleaving between pairs is enabled. If so, the algorithm first checks whether the address  
10 is within the interleaved memory range, as shown in step 904. If the cache line address is above the interleave range, then the present invention stacks on the larger MSU\_PAIR, as shown in step 910. Otherwise, flow continues to step 906 where it is determined which MSU\_PAIR should be selected between the plurality of MSU\_PAIRs. In a preferred embodiment, the low order cache line address bit, TEMP\_ADDR [0] is used to select the  
15 MSU\_PAIR.

If interleaving between pairs is not enabled, then the present invention stacks cache line addresses. In a preferred embodiment, the present invention begins stacking the cache line addresses into MSU\_PAIR0. Once MSU\_PAIR0 (i.e., MSU\_Pair 222) is full, then  
20 stacking proceeds to MSU\_PAIR1 (i.e., MSU\_Pair 224). Stacking continues until the highest MSU\_PAIR is full. This is shown generally at step 912.

Flow then continues to step 908 (from either block 906, 910 and 912) where the cache line address is readjusted. The adjustment depends upon whether the interleaving or  
25 stacking is chosen. In the case of interleaving, the cache line address (TEMP\_ADDR) is readjusted by shifting the address to the right by one location and zero-filling the most significant address bit. In the case of stacking, the cache line address either remains the same or is set equal to TEMP\_ADDR - SMALLEST\_PAIR\_SZ, as evident by a review of the pseudo-code.

30

- 48 -

Once an MSU\_PAIR is selected for stacking, the present invention proceeds to stage 920. This stage of the algorithm has an input TEMP\_ADDR, which may have been adjusted by step 908. Stage 920 uses the following registers: PAIR0\_MODE, PAIR0\_SMALLEST\_MSU\_SZ, PAIR0\_SEL. The outputs from stage 920 are  
5 TEMP\_ADDR, which is the cache line address after any required adjustments, and RCVING\_MSU, which indicates which MSU will receive the cache line address. At initialization, PAIR0\_TOP\_OF\_INTLV\_RANGE is the address value where no more memory is left for interleaving between MSUs of MSU\_PAIR0. PAIR1\_TOP\_OF\_INTLV\_RANGE is the address value where no more memory is left for  
10 interleaving between MSUs of MSU\_PAIR1.

If Stage 900 selected MSU\_Pair0, then stage 920 determines whether RCVING\_PAIR equals MSU0 or MSU1. Similarly, if stage 900 selected MSU\_Pair1, then stage 920 determines whether RCVING\_PAIR equals MSU2 or MSU3. For the sake  
15 of brevity, only a selection between MSU0 and MSU1 will be described.

Step 924 determines whether interleaving between the multiple MSUs of an MSU\_PAIR is enabled. If interleaving is enabled, the algorithm first checks whether the cache line address is within the interleaved memory range, as shown in step 926. If the  
20 cache line address is within the interleaved memory range, the low order cache line address bit is used to select the appropriate MSU, as shown in step 928. Next, the cache line address is readjusted by shifting the cache line address bits to the right by one location and zero-filling the most significant address bit, as shown in step 930.

25 If, on the other hand, the cache line address is above the interleave memory range, then the algorithm stacks onto the larger MSU, as shown in step 932. Flow then proceeds to the step 930 where the address is adjusted for stacking by setting TEMP\_ADDR to TEMP\_ADDR - PAIR0\_SMALLEST\_MSU\_SZ.

30 If interleaving between MSUs of the MSU\_PAIR0 is not enabled, the

- 49 -

present invention stacks MSU0 first and then stacks the remainder into MSU1, as shown in step 934. Once again, the address is adjusted in step 930 based on whether the low or high MSU is used first. When the low MSU is used first, TEMP\_ADDR remains unchanged. When the high MSU is used first, TEMP\_ADDR is set to TEMP\_ADDR - PAIR0\_SMALLEST\_MSU\_SZ.

As stated above, a similar procedure is followed for selecting between MSU2 and MSU3 in MSU\_PAIR1.

Finally, as shown in step 940, MSU\_ADDR [29:0] is assigned to the adjusted TEMP\_ADDR [29:0] and the RCVING\_PAIR is concatenated with the RCVING\_MSU indicators to form MSU\_SEL [1:0]. This completes the forward address translation algorithm.

Shown in Appendix B is pseudo-code for the reversed translation algorithm. The reverse address translation function resides only in the MSU controller (not shown).

Reference to Figure 6 will be made to demonstrate an example of the forward address translation algorithm. Figure 6 illustrates a main memory 600 having two MSU\_PAIRs 610, 640. MSU\_Pair 610 has two MSUs 620, 630, whereas MSU\_Pair 640 has a single MSU 650. MSU 620 has one 128 megabyte memory bank 1020, MSU 630 has two 128 megabyte banks 1030 (or 256 megabytes of memory space), and MSU 650 has four 128 megabyte banks 1040 (or 512 megabytes of memory space). The top of MSU 620 is 80.0000<sub>H</sub>. This means that 80.0000<sub>H</sub> is the address location where there is no more memory left for interleaving. The top of MSU 630 is 100.0000<sub>H</sub>. Thus, MSU\_Pair 610 has a pair size of 180.0000<sub>H</sub>. The top of MSU 650 is 200.0000<sub>H</sub>. Thus, MSU\_Pair 610 has a pair size of 200.0000<sub>H</sub>. Note that MSU\_Pair 640 is treated conceptually as a pair of MSUs even though it includes only a single MSU 650.

Suppose there are four cache line addresses 0.0000.0000<sub>H</sub>, 0.0000.0040<sub>H</sub>, 0.0000.0080<sub>H</sub>, and 0.0000.00C0<sub>H</sub>, respectively representing four memory references from

- 50 -

four operating systems following any address relocation performed. For this example, main memory is configured as shown in Figure 6. Note that this is not the most efficient memory configuration for this number of memory banks.

5       The register setup for this example is as follows: PAIR\_MODE equals 0 (Interleave), PAIR0\_MODE equals 0 (Interleave), PAIR1\_MODE equals 1 (Stack), SMALLEST\_PAIR\_SZ equals 003<sub>H</sub>, PAIR0\_SMALLEST\_MSU\_SZ equals 001<sub>H</sub>, PAIR1\_SMALLEST\_MSU\_SZ equals 004<sub>H</sub>, PAIR\_SEL equals 1, PAIR0\_SEL equals 1, PAIR\_SEL equals 0. The above setup represents the IIS option of translation.

10

Using these register settings and presenting the first address to the algorithm yields the following results:

Initialization for both phases:

PROCESSOR\_ADDR[35:0] = 000000000<sub>H</sub>

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

TOP\_OF\_INTRLV\_RANGE = 003<sub>H</sub>

PAIR0\_TOP\_OF\_INTRLV\_RANGE = 002<sub>H</sub>

PAIR1\_TOP\_OF\_INTRLV\_RANGE = 004<sub>H</sub>

15       the MSU\_Pair selection phase:

In

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

Results:

RCVING\_MSU = 0 (MSU\_PAIR0)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

the MSU# selection phase:

- 51 -

In

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

Results:

RCVING\_MSU = 0 (MSU#0)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

the final results:

MSU\_ADDR[29:0] = 000000000<sub>H</sub>

MSU\_SEL[1:0] = 00 (MSU#0 of MSU\_PAIR0)

5

Processing the second address

Initialization:

PROCESSOR\_ADDR[35: 000000040<sub>H</sub>  
0] =TEMP\_ADDR[29:0] = 00000001<sub>H</sub>

RCVING\_PAIR = 1 (MSU\_PAIR1)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

RCVING\_MSU = 0 (MSU#2)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

the final results:

MSU\_ADDR[29:0] = 00000000<sub>H</sub>MSU\_SEL[1:0] = 10 (MSU#2 OF  
MSU\_PAIR1)

10

- 52 -

The third address yields:

Initialization:

PROCESSOR\_ADDR[35: 000000080<sub>H</sub>  
0] =

TEMP\_ADDR[29:0] = 00000002<sub>H</sub>

RVCING\_PAIR = 1 (MSU\_PAIR1)

TEMP\_ADDR[29:0] = 00000001<sub>H</sub>

RCVING\_MSU = 0 (MSU#2)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

Final results:

MSU\_ADDR[29:0] = 00000000<sub>H</sub>

MSU\_SEL[1:0] = 01(MSU#1 OF  
MSU\_PAIR0)

5

While the fourth address yields the final results:

Initialization:

PROCESSOR\_ADDR[35: 0000000C0<sub>H</sub>  
0] =

TEMP\_ADDR[29:0] = 00000003<sub>H</sub>

RVCING\_PAIR = 1 (MSU\_PAIR1)

TEMP\_ADDR[29:0] = 00000001<sub>H</sub>

RCVING\_MSU = 0 (MSU#2)

TEMP\_ADDR[29:0] = 00000000<sub>H</sub>

- 53 -

Final results:

MSU\_ADDR[29:0] = 00000000<sub>H</sub>

MSU\_SEL[1:0] = 01(MSU#2 OF  
MSU\_PAIR1)

Figure 7 shows the result of this example.

5

It should be understood that embodiments of the present invention can be implemented in hardware, software or a combination thereof. In such embodiments, the various components and steps may be implemented in hardware and/or software to perform the functions of the present invention. Any presently available or future developed computer software language and/or hardware components can be employed in such  
10 embodiments of the present invention. In particular, the pseudo-code discussed and provided above and in the appendixes below can be especially useful for creating the software embodiments.

15

### ***C. Initialization at Boot Time***

In an exemplary embodiment, partitioning of the computer system 200, including the processing modules and the memory 160, in accordance with the present invention, is performed at boot time. Exemplary processes for partitioning, mapping memory and  
20 setting up interleaving, are described below. These initialization operations can be performed by a Basic Input/Output System (BIOS) and a Management Interface Processor (MIP) at boot time via an MIP high-speed scan interface. The MIP is a hardware interface portion of a management application platform (MAP) for performing initialization and error recovery for the computer system 200. In an exemplary embodiment, the MIP high-  
25 speed scan interface complies with IEEE TAP Linker Specification 1149.1.

- 54 -

As used herein, the term "partition" is sometimes used in place of window. As used herein, these two terms are synonymous, and indicate a part of the system that is controlled by one instance of an operating system.

5           The manner in which the partitioning will be accomplished at boot time can be pre-determined by a system administrator and entered into a database that resides on MAP. Partitioning information identifies system resources which are to be allocated to a particular window, which type of operating system will be loaded within the window, and whether and how two partitions will communicate via shared memory. In the exemplary  
10           embodiment of Figure 2, partitioning preferably occurs on sub-Pod and direct I/O bridge (DIB) boundaries.

          Generally each operating system has certain hardware requirements. For example, off-the-shelf, open architecture operating systems, such as Windows NT and Unixware  
15           (available from The Santa Cruz Operation, Inc.), require a disk controller (SCSI fiber channel, etc), VGA controller, compatibility PCI board, and compatibility peripherals (CD-ROM, tape, and disk). The appropriate hardware should be resident on the system, and the system should be partitioned in a manner that ensures these requirements are met. This should be taken into account when entering the partitioning information into the database  
20           on the MAP.

Referring to Figure 13, a process flowchart is provided to illustrate an exemplary initialization process:

25           Processing begins at step 1310, where the MIP loads the BIOS into main memory.

          In step 1312, the MIP loads the BIOS configuration data area in main memory. This information partially reflects what was stored in the configuration database.

30           In step 1314, the MIP releases each sub-Pod from reset one at a time. Preferably, the sub-Pods arbitrate and one sub-Pod becomes the BIOS sub-Pod (BSP). Within the



- 55 -

BSP, one processor becomes the master, and this processor executes the BIOS code. Throughout the remainder of this specification, the processor that runs the BIOS can be referred to as the BSP. The BSP performs a number of functions, as described below.

5           In step 1316, the BSP initializes each PCI Bus. The BSP gains access to each PCI Bus in the system through a path that extends from the Crossbar Interconnect in the BSP's sub-Pod, to the MSU, through another Crossbar Interconnect on another sub-Pod, and finally through an interface to the DIBs. The BSP can access the DIBs associated with its own sub-Pod without accessing the MSU.

10

          In step 1318, the BSP reads configuration data, which was loaded into main memory in step 1312, above, to determine which DIBs are in which partition. The BSP writes a Partition ID (PID) to a "DIBs in the Partition Register" in each Compatibility DIB by using the path described above. The PID is used when a message is received by a DIB  
15   during normal system operations. The message is only processed if the DIB has the same PID as the message. The PID allows all units in a partition running under the same operating system to talk to one another, and is also used to send messages through shared memory.

20

          In optional step 1320, the BSP calculates the size of the high memory hole and low memory hole by reading PCI registers in each of the PCI cards to determine I/O and memory requirements for each PCI cards. Overlaying I/O space with main memory is required by the Intel Multi-Processor Specification, and by the fact that some off-the-shelf PCI cards can not recognize addresses above 64 gigabytes.

25

          In step 1322, the BIOS informs the MIP of the amount of memory-mapped I/O space that is required by each PCI card. This is done via a BIOS-to-MIP interrupt and associated mailbox. The MIP already is aware of the size of main memory, and the amount of memory that is to be shared between operating systems, because this  
30   information is included in the configuration database associated with the MIP. Therefore,

- 56 -

after the MIP is informed as to the amount of I/O space required, the MIP can calculate the following information using Tcl scripts:

- a. Location of the high and low memory holes
- b. Location of reclamation area
- c. Location of shared memory

Tcl is an industry-standard simulation language that is used by the hardware designers to write simulation scripts. The simulation scripts are also ported to the MIP to accomplish hardware initialization.

In step 1324, the MIP uses the memory addresses calculated above along with data located in the configuration database to set up registers in the sub-Pods (TCT), crossbar interconnect (TCM), and memory storage unit (MSU). Initialization of the TCM sets up the partitioning and address translation for the DIBs and memory address translation registers for the DIB. These constants can be used for interleave functions and memory reclamation.

In an exemplary embodiment, there are at least two sets of registers in each TCM, one for each DIB. These include range registers and broadcast registers.

Range registers for the DIB contain the legal memory range for each DIB, according to the partition definition. Interfaces within the TCM are enabled/disabled according to partition definitions.

A TCT Info Register is initialized with, among other things, the Partition ID, which identifies the partition. This is used to determine if a particular sub-Pod should operate on messages. Messages having the same Partition ID as in this register will be received.

- 57 -

Broadcast registers contain the Partition ID, and are used for broadcast messages throughout a partition. A broadcast message is tagged with the Partition ID as identified in this register.

5       Agent tables are loaded with the Partition ID, and are used to validate interrupts to the processors of a particular window.

      In the DIB, range registers for the PCI Cards contain address ranges for memory-mapped spaces for each PCI bus. A Partition ID register contains the Partition ID so that  
10   only messages for that DIB are received.

      In the MSU, MSU\_PairA/PairB configuration registers set up interleave between banks of MSU. The MIP initializes the Memory Address Translation Registers (see Tables E and F above) to set up interleave operations. These interleave operations are specified by  
15   the user prior to initialization.

      The MIP uses the length of the memory-mapped I/O space as received from the BIOS to calculate the locations of the memory-mapped I/O space, the shared memory start address, the reclamation start address, and new top of memory. The MIP communicates  
20   these start addresses back to the BIOS using the MIP-to-BIOS interrupt and associated mailbox in main memory. The MIP further uses this information in conjunction with user-specified configuration data to initialize the Configuration Register (Table A, above), and the Translation and Reclamation Register (Table D, above). The initialization data stored in these registers and the Memory Address Translation Registers (Tables E and F, above) is  
25   required by the address translation logic to perform the windowing, reclamation, and address translation functions. As discussed above, copies of these registers and the associated logic are located in each of the TCTs 270 (for memory requests from a processor 240), and are also located in each of the TCMs 285 (for memory requests from an I/O processor via a DIB). The MIP further initializes range registers for the processors  
30   with valid address ranges for the memory-mapped space for each DIB, I/O port, APIC memory-mapped space, and memory address space.

- 58 -

The BIOS uses this information to set up a configuration table in memory for each partition/operating system. This information communicates the location of shared memory to each partition. The configuration table could be of any user-defined format. In an exemplary embodiment, an MP Configuration Table, as defined in a MultiProcessor Specification available from the Intel Corporation, is used. The field called "OEM Table Pointer" within the MP Configuration Table is used to point to a user-defined area that will include the location and length of the shared memory area. Unixware and NT drivers use this information for memory allocation purposes, and to determined queue locations.

10

The BIOS further sets up registers in selected ones of the processors. The BIOS sets up these registers because the MIP does not have access to them. In an exemplary embodiment, this is just done for Intel processors, and involves writing registers within each of the processors to indicate, for example, a top of memory register (TOMR) in each processor that communicates to an operating systems where the top of memory is. The operating system is not allowed to attempt to access memory above the TOMR value.

15

Registers can also include memory type range registers (MTRR) that communicate to processors which type of memory exists within the various memory ranges (e.g., mapped I/O, APIC interrupt space, main memory, etc.). MTRRs are used to tell processors how to handle memory accesses. For example, processor read operations to a memory range that is designated as memory-mapped I/O space are not cached in the processor's cache. Processors running an instance of operating system should have the same value loaded into their respective MTRR.

20

25

In step 1326, after performing any additional initialization functions, the BIOS reads the boot sector for each operating system into the appropriate location in memory as determined by information in the configuration database.

30

In step 1328, the BIOS issues an interrupt to one of the processors in each partition, and those processors begin loading the associated operating system from a designated I/O

- 59 -

device. When this is completed, the operating system assumes control of the resources in its window. This completes the BIOS to operating system transition and processing.

### III. Methods for Managing the Global Shared Memory (Inter-Partition Communications)

5

The global shared memory approach described above can provide a private memory space for each partition, plus a shared memory area that all of partitions can access. The shared memory area can include one or more read-only areas. Partitions, including their operating systems and other clients running on the partitions, can communicate with one another through the shared memory.

The shared memory area can be managed by, for example, a portion of the operating system running on a partition, or by other software and/or hardware that may reside on a partition. The shared memory area can be managed by different operating systems, including, but not limited to, Windows NT, commercially available from Microsoft Corp., UNIXWARE, commercially available from The Santa Cruz Operation, Inc. (SCO), Master Control Program (MCP), which is an operating system adapted for UNISYS Clearpath HMP NX computer systems, which supercede the A-Series family of computer systems, both of which are commercially available from Unisys Corp., or OS 2200, which is an operating system adapted for UNISYS Clearpath HMP IX computer systems.

Alternative embodiments are described below for managing a shared memory area in accordance with the present invention. The embodiments are described herein for purposes of illustration, and not limitation. Other embodiments (including equivalents, extensions, variations, deviations, et cetera, of the embodiments described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. The invention is intended and adapted to include such alternate embodiments.

- 60 -

### **A. Polling For Inter-Partition Communications**

In one embodiment, each operating system executing in its own partition (e.g., one or more Pods or sub-Pods) on the computer system is associated with, or allocated, a portion of shared memory 160. Operating systems can write to, and read from, their associated portions of shared memory but cannot write to portions of memory associated with other operating systems. All operating systems can, however, read from the entire shared memory.

Preferably, each partition or operating system is assigned an exclusive memory window (sometimes hereinafter also referred to as its "local memory space") dedicated to the partition or operating system. When an operating system or an application associated with the operating system sends a message to another operating system, or to an application associated with the operating system, the sending entity builds the message in a buffer in its local memory space in the same manner as would occur if the message were being built to be transferred via a network. The sending entity then copies part, or all, of the message, into its allocated part of shared memory 160.

The target partition/operating system, which can read from, but which cannot write to, the sending operating systems' associated portion of shared main memory 160, detects that a new message is available, and copies the message from shared main memory into its own local memory (*i.e.*, its exclusive memory window).

In an exemplary embodiment, code and most data structures for an operating system reside in the local memory space for the operating system. Certain new data structures preferably reside within the shared memory 160.

In an exemplary embodiment, two types of data structures are used to facilitate communication between partitions or operating systems. The first type includes message storage structures that store the message data, and which are built in output message buffers. The second type includes queue structures that are stored within a message queue

- 61 -

area and which contain pointers to message data stored in an associated output message buffer. Preferably, these two types of data structures are stored in shared main memory 160 while other code and data constructs utilized by the various operating systems and associated application programs reside in associated local memory spaces. This protects  
5 system integrity.

Figure 14 illustrates a portion of shared memory 160, including an output message buffer pool area 1402 and a message queue area 1414. Generally, an output message buffer pool area 1402 is associated with each partition. Buffers 1410 are allocated for a  
10 message and pointed to by a queue entity, or multiple queue entities, when a message is broadcast

Generally, all partitions have read access to all output message buffer pool areas 1402. But each partition has write access only to buffers 1410 in its associated output  
15 message buffer pool area 1402.

Message queue area 1414 is divided into  $n$  node output queues 1412, each of which is dedicated to a different partition. Although all partitions have read access to the entire message queue area 1414, a partition can only modify its associated node output queue  
20 1412. This access control, which can be enforced within hardware, renders hardware locks unnecessary, thereby simplifying recovery and checkout operations.

Figure 15A illustrates an exemplary embodiment of message queue area 1414 is illustrated with eight node output queues 1412. Node output queue 1412a is illustrated  
25 including a node-to-node queue 1510 for each partition. As used herein, the term "node" is equivalent to the term "partition."

Figures 16A and 16B illustrate exemplary information contained in a node output queue 1412. The first sixteen words of an exemplary node output queue 1412 includes  
30 control information for the associated node including node operating system type

- 62 -

(Node\_OS\_ID) 1610, node media access control (MAC) address 1612, and various reset flags (e.g., Reset\_OK) used during recovery, as is discussed below.

The control information further includes eight Dequeued\_offset fields, each of which stores an offset into a respective different node's node output queue, and indicates which is the next message to receive from that respective different node, as will be explained below.

In the exemplary embodiment of Figures 16A and 16B, node-to-node queues 1510 follow the first sixteen words of control information. Each node-to-node queue 1510 is used by the associated operating system to send messages to the named different node. For example, node 0 to node 1 queue 1510a is used by node 0 to send messages to node 1. For simplicity, a node-to-node queue 1510 can be provided for each node to send a message to itself.

In Figures 16A and 16B, the first word in each node-to-node queues 1510 includes control information including a "Need\_Reset" flag and an "Enqueue\_offset". The Need\_Reset is used in conjunction with a selected one of the Reset\_OK flags when a sending node wants to reset one of the node-to-node queues. The "Enqueue\_offset" contains a number between 1 and 511, for example, and is used to point to the next available entry in the respective node-to-node queue 1510. The use of this field is explained further below. Each of the remaining words (e.g., 511 words) of the node-to-node queue 1510 includes an offset pointer that points to an associated message data structure 1416 in an associated output message buffer 1410

In a preferred embodiment, the offset is the number of 64-bit words from the start of the respective node's output message buffer 1410. The pointer should be an offset from some base address, not a real or virtual address. The pointer should not be based on a virtual address because, when the nodes are heterogeneous nodes, they may not have a common virtual address translation. The pointer should not be based on a real address because, as a result of the address translation scheme described above, real addresses used by one node generally do not coincide with real address used by another node.



- 63 -

In an exemplary embodiment, pointers are offsets from an address that each node or operating system can calculate from information received from the Management Application Platform (MAP), described above, during node initialization.

5

Each of the eight node-to-node queues 1510 within a node output queue 1412 can be, for example, 512 words long, as illustrated in Figures 16A and 16B, so that each node output queue 1412 is  $16 + 8(512)$  words long.

10

This queue depth helps to ensure that an associated queue will not be full when a message is available to be transferred to shared memory. The queue depth may be specified by the manager application platform (MAP) during initialization. As mentioned above, the MAP is a support system for performing initialization and error recovery on the computer system 200.

15

To add flexibility, the MAP can be designed to indicate the queue capacity at initialization time. This data may be added as an entry into each of the configuration tables, which are data structures provided by MAP for each operating system instance in the system to inform the respective operating system of necessary system parameters such as the location of shared main memory.

20

Figure 17 illustrates an exemplary embodiment of a message data structure 1416. Each message data structure 1416 is preferably located at an offset of 0 within an associated output message buffer 1410 and includes a header area 1710 and a message data area 1712. The header area 1710 is illustrated as occupying words 0-n, and includes the buffer length, header length, and count information. The count information is preferably included for writing messages by a 2200 operating system (*i.e.*, an operating system adapted for a 2200 style processor commercially available from Unisys Corporation) because messages written to memory by the 2200 operating system will not occupy contiguous memory locations. When nodes running the 2200 operating system record message data in shared memory, each 64-bit main memory word will store, at most, 32 bits

25

30

- 64 -

of data which will be located in the least significant bits of each 64-bit main memory word. Some words may store fewer bits if messages do not start or end on a word boundary. Therefore, the first Byte Skip Count indicates the number of bytes that should be skipped between a protocol header and the message data. The Byte Transfer Count indicates the  
5 byte length of an associated valid message field. The sum of the Byte Skip Counts and Byte Transfer Counts should be less than or equal to (length of the buffer - length of the header) \* 4.

In an Ethernet environment, the maximum message segment size is 1500 bytes or  
10 375 64-bit words for the message. In an embodiment, the present invention includes a network input/output processing architecture (NIOP), which is a message handler developed by Unisys Corporation, as described in U.S. Patent Number 5,659,794, assigned to Unisys, which allows for 50 separate data streams to be combined into one message segment to be sent over a network. Therefore, an output message buffer size of 427 words  
15 would allow a 2200 operating system to continue to perform in the shared memory environment of the present invention as it does for an Ethernet LAN environment. Given a queue depth of 511 and a buffer size of 427 words, a node buffer pool size of  $(511 * 427 * 8) // 4096 = 1,748,992$  words. The total shared memory needed per shared memory environment is then  $(65,536 + 1,748,992 * 8) // 4096 = 14,057,472$  words.

The use of these data structures can be explained by example. Assume that a first operating system OS1 wants to send a message to a second operating system OS2. Further assuming that the OS1 to OS2 node output queue 1412 is not full, OS1 obtains an available message data structure (i.e., buffer) 1416a within the OS1 output message buffer area  
25 1410a. The buffer 1410a is preferably identified by an address offset pointer as discussed above. OS1 builds a protocol header 1710 for the message, transfers the header 1710 and message 1712 from the local main storage of OS2 into this available message buffer 1416a. OS1 then increments the contents of an Enqueued\_offset within the OS1 to OS2 queue 1510a to point to the next available entry in the OS1 to OS2 queue 1510a. OS1 copies the  
30 offset pointer which points to the message data structure (i.e., buffer) 1416a into this next

- 65 -

available entry. In a preferred embodiment, the Enqueued\_offset maintained as a circular queue.

OS2 polls to determine if a message is available from OS1. This is done by  
5 comparing the contents of an appropriate Dequeued\_offset for OS2, stored in the control area of OS2's node output queue 1412a, to the appropriate Enqueued\_offset stored in the OS1 to OS2 output queue of OS1's node output queue 1412b. In a preferred embodiment, the Dequeued\_offset is maintained as a circular queue.

10 Each of the eight Dequeued\_offsets (in the exemplary embodiment) stores a value between 1 and 511 which points to an entry within a corresponding sending one of the node's node output queues 1412. For example, the Dequeued\_offset stored within word 8 of OS2's output queue stores an offset value which points into the "Node 0 to Node 1 Queue" within OS1's node output queue 1412a. Similarly, the Dequeued\_offset stored  
15 within word 15 of OS2's node output queue 1412 stores an offset value which points into the "Node 7 to Node 1 Queue". As noted above, the data structures include a node output queue 1412 and associated Dequeued\_offset which allows each node or operating system to send a message to itself, e.g., OS1 to OS1 node output queue.

20 In the current example, the Dequeued\_offset field within word 8 of the OS2 node output queue 1412 is compared to the Enqueued\_offset field within the OS1 to OS2 queue. If the two offset entries are the same, the queue is empty. When the Enqueued\_offset is different than the Dequeued\_offset, one or more entries exist on the OS1 to OS2 queue.

25 After OS1 determines a message is available, it uses the contents of the Dequeued\_offset to retrieve the message and then increments the Dequeued\_offset. The message offset pointer is used to retrieve the message, which is copied into local storage.

A sending node or operating system can use a mechanism similar to the above-described polling mechanism to determine whether a queue is full prior to adding an entry  
30 to the appropriate queue. That is, the Dequeued\_offset within the recipient's queue is compared to the appropriate Enqueued\_offset within the sending node's output queue. If

- 66 -

the contents of the Enqueued\_offset is the same as the contents of the Dequeued\_offset, the queue is full and no message may be added at that time. Enqueued\_offsets and Dequeued\_offsets conforms to the assumption that all operating systems may read all other operating systems' queue areas, but an operating system may only modify its own queue  
5 area.

In a virtual memory system, code and/or data structures can be transferred, or "paged", out of main memory to mass storage under the direction of an operating system to make additional room within the main memory. In an exemplary embodiment of the  
10 present invention, paging out is allowed for code and/or data structures stored within a local memory area, but is not allowed for data structures residing in shared memory 160. This restriction ensures that operating systems that use shared memory space 160, can make assumptions about the location and content of the data structures stored within the shared memory space 160.

15

In an exemplary embodiment, 2200 operating system applications communicate with Intel-based applications (e.g., applications written for Windows NT on an Intel platform) wherein the only substantial operating system involvement is managing the shared memory (e.g., requesting initialization of the message queues). In this exemplary  
20 embodiment, the 2200 operating system does not request services or perform services for the Intel nodes. Instead, services are performed through application-to-application requests. One skilled in the relevant art(s) will recognize that the 2200 operating system could, alternatively, be altered to directly request services of the Intel node.

25

In an exemplary embodiment, the global shared memory mechanism allows communication to occur between 2200 operating system application programs and NT and/or Unix application programs. It can also be used to facilitate communication between applications running under the MCP operating system and applications running under a NT and/or Unix operating system, and can be used for communication between operating  
30 systems. Similarly, it can be used to facilitate communications between applications running under an associated different instance of an NT operating system and for

- 67 -

communications between applications running under an associated different instance of a Unix operating system. The shared memory mechanism can be used to facilitate communication between 2200 and MCP operating systems.

5 In an exemplary embodiment, messages written to shared main memory are typically ASCII characters, but can also include positive integers such as one, two or four-byte positive integers, and bit information. 2200 operating systems, which operate on 36-bit words, represent ASCII characters as 8 bits within a 9-bit byte. Intel platforms, which use IA 32 or IA 64 architecture and operate on 32-bit or 64-bit words, respectively,  
10 represent ASCII characters as 8 bits within an 8-bit byte. Therefore, data written to, or read from, shared memory should undergo a conversion process. This conversion can be performed by 2200 operating system hardware instructions. A 2200 style processor uses a Block Transfer Pack (BTP) instruction to pack ASCII data from 9-bit to 8-bit bytes, and to zero fill the most significant 32 bits within the 64-bit words of the main memory.

15 Typically, applications running on Intel platforms expect that message data is included within contiguous bytes. Since the 2200 operating system Block Transfer Pack (BTP) instruction does not enter the message data in contiguous bytes within the shared main memory (four bytes within a word are usually unused), device drivers operating on  
20 Intel platforms must move the message data into contiguous bytes within local main memory before the message can be processed. Similarly, when a 2200 style processor receives a message, it uses a Block Transfer Unpack (BTU) instruction to unpack ASCII data from shared main memory and move it to associated local memory. The Block Transfer Pack and Block Transfer Unpack instructions also perform big-endian/ little-  
25 endian conversion. Examples of data movement into and out of shared memory 414 for a 2200 to Intel message, an Intel to 2200 message, and an Intel to Intel message are provided below.

30 Preferably, the global shared memory communication mechanism is as transparent as possible to the software running on the system so that software changes are minimized and so that the system is as compatible as possible with various open-system standards.

- 68 -

For example, in accordance with an aspect of the present invention, the system can be made to appear from the upper layers of the software as though communication-by-wire has been maintained (see Section IV.B. below). In an exemplary embodiment, the system employs an Ethernet protocol. One skilled in the relevant art(s) will recognize that other  
5 protocols, such as an ATM protocol can be used.

For NT/UNIX nodes, a Shared Memory interface is preferably visible within a NIC device driver, which exists at the LLC/MAC level of the Open Standards Interconnection (OSI) communications model. LLC/MAC are 2 sub-layers of the OSI level 2  
10 communications model. LLC can be an interface between layers 2 and 3. MAC is an IEEE sub-layer that deals with various LANs such as Ethernet, Token Ring, Token Bus, etc.

In 2200 operating system nodes, this visibility also occurs at the LLC/MAC level.  
15 This design choice also makes it easy to allow some partitions to communicate through shared memory while other partitions maintain communication via a wire. The two types of communication are viewed as the same from the upper layers of the software.

Since the Ethernet protocol imposes a limit of 1500 bytes per transmission, a large  
20 message may have to be divided into segments and transferred during multiple message transfer operations.

Ethernet has a 1500 byte limit on the amount of data in one transmission. Thus, where an Ethernet connection is replaced with shared memory, 1500 bytes becomes the  
25 limit on how much data can be placed in a buffer that is queued for output to another node. As with all communications protocols, any size message can be sent, but it may have to be sent in a number of separate transmissions (buffers).

A 2200 style processor can transfer message data into shared memory using the  
30 Block Transfer Pack instruction discussed above.

### ***B. Interrupt-Driven Shared Memory Communications***

An interrupt-driven shared memory management implementation is now described, as an alternate embodiment, including a description of how the shared memory area, or region, is to be accessed and managed in accordance with this alternate embodiment. In this embodiment, management of the shared memory window is performed by program code embodied in Core Services software that resides on each partition. The Core Services software on each partition provides an application programming interface (API) that a client running in that partition can call to request certain shared memory services, such as, for example, communicating with a client on another partition via the shared memory window. As used herein and in the claims, a "client" can be the operating system, a device driver, an application program, or any other software or program code executing on a partition that requires the use of the shared memory window. Also as used herein and in the claims, the term "a communication" may mean a Signal (described hereinafter), a message in the form of data (which may or may not be stored in an allocated buffer within the shared memory window), or any other form of information or data to be communicated between partitions for any purpose. Unlike in the previous embodiment, in which a polling technique is employed to determine whether a communication is to be transferred between partitions, this embodiment employs an inter-processor interrupt mechanism to communicate between partitions, as described more fully below.

As with the previous embodiment, this embodiment can be used to facilitate communications between partitions running under the control of different operating systems (*e.g.* Unisys MCP, Unisys OS 2200, Windows NT, Unix, etc.) or partitions running under the control of different instances of a same operating system.

#### **1. Shared Memory Layout**

Figure 19 illustrates the layout of the shared memory window in accordance with this alternate embodiment. As shown, a control structure 1900 resides at the base of the shared memory window, followed by the remainder of the shared memory

- 70 -

window, 1916 which is broken into separate pages. In the present embodiment, each page comprises 4K bytes, however, the size may be different in other embodiments. Each page can be in-use, available, or out-of-use. As described hereinafter, a client can request that a portion of the shared memory window be allocated to it, for example, to  
5 define a buffer, and the Core Services software then allocates the required number of pages to satisfy that request.

The shared memory control structure 1900 comprises a header 1910, an allocation structure 1912, and a plurality of partition input queues with an associated header 1914. Information in the control structure is private. Direct access to this  
10 information is not provided to clients of the Core Services software. Instead, the Core Services software API provides calls that return client-related information to a client via procedural parameters. In the present embodiment, words in the control structure include 64 bits, where the upper 32 bits are 0's to allow for the different size words used by different processor architectures.

15

## 2. Free Page List

In the present embodiment, in order to keep track of available shared memory pages, *i.e.*, those that are not already in-use, the available pages are linked through pointers in the first word of each page to form a linked-list of available pages. The  
20 linked-list of available pages is referred to herein as the Free Page List. The control structure 1900 provides a pointer to the first page of the linked list (*i.e.*, the start of the Free Page List).

## 3. Client Directory Table

25 The Core Services software allocates one or more pages of the shared memory window to store a Client Directory Table (not shown). The Client Directory Table is a registry of the clients on each partition that are using the shared memory window. More specifically, in the present embodiment, each client of the Core Services software on a given partition must register with the Core Services software as a member of a  
30 Client Group. Two clients on the same partition cannot be members of the same Client



- 71 -

Group; if there are multiple clients of the Core Services software on a partition, each must register as a member of a different Client Group. Each Client Group has an associated name (Client Group Name) and identifier (Client Group ID). The Client Directory Table contains an entry for each Client Group that specifies the Client Group Name and lists each partition that has a client registered as a member of that group. When a client registers with the Core Services software as a member of a particular Client Group, the Core Services software returns the Client Group ID to the client. The Client Group ID is used to identify the sending and receiving clients when messages are passed via the shared memory window, as described hereinafter.

#### 4. Shared Memory Page Types

The Core Services software may allocate one or more pages of the shared memory, either for its own use or on behalf of a client request to allocate some portion of shared memory. In the present embodiment, four different page types are defined.

##### a. Type 1 Memory Pages

Type 1 memory pages, in the present embodiment, can only be allocated for use by the Core Services software on a partition; there are no interfaces to allow a client to request allocation of a Type 1 page. As one example, the Client Directory Table described above is stored in one or more Type 1 pages allocated by the Core Services software. When the Core Services software allocates a Type 1 memory page, a Core Services header is created at the beginning of the page. Figure 32A illustrates the contents of the Core Services header for Type 1 pages, in accordance with the present embodiment.

The first field (Partition Ownership Mask) is used to store an indication of which partitions have access rights to the page. Specifically, the Partition Ownership Mask contains eight bits, one for each possible partition in the computer system. Each partition that has ownership rights to the page will have its corresponding bit in the Partition Ownership Mask set. In the case of the Client Directory Table, for example,

- 72 -

each partition that requires access to the table will have its bit of the Partition Ownership Mask set in each page containing all or part of the table.

Although in the present embodiment, there are no interfaces to allow a client to request allocation of Type 1 pages, to accommodate future embodiments in which it may be desirable to allow clients to request allocation of Type 1 pages, the Core Services header in a Type 1 page further contains a Client Group ID field. This field would be used to contain the Client Group ID of the clients that have ownership rights to the page. In the present embodiment, however, this field is not used.

The DeallocationLock field is used to coordinate changes in the ownership of the page. This field is part of a broader lock mechanism of the present invention, implemented throughout the Core Services software, that allows different partitions to lock access to the various structures, pages, and tables of the shared memory window, as needed, and in a consistent manner, to ensure that only one partition is capable of modifying any given structure, page, or table at a time (*i.e.*, to synchronize access to these structures).

The DeallocationLock field, as well as all other lock fields described hereinafter, consists of two 64-bit words, designated Word 0 and Word 1. Word 0 defines a Lock Status Word, and Word 1 defines an Owner Word. The low order bit of Word 0 defines an "in use" bit. Setting this bit indicates a locked status. Word 1 is used to store the Partition ID of the partition that acquires the lock, enabling the owner of the lock to be determined.

Most operating systems and the processors on which they execute, provide a method by which the operating system and clients executing under those operating systems can acquire a lock to a given data structure. The lock field format used herein is compatible with a number of operating systems, including, for example, Windows NT, UnixWare, and the Unisys MCP. The Core Services on a given partition must be tailored to the operating system and processor architecture of that partition.

30

- 73 -

In accordance with an important feature of the lock mechanism of the present invention, when a Type 1 memory page is first allocated, the allocating partition must acquire a system wide lock (a field of the Allocation Structure described hereinafter) in order to lock access to the page during allocation. However, when ownership of one or  
5 more allocated pages is extended or transferred to other partitions, only a lock to the pages involved must be acquired. The DeallocationLock field in these pages is used for this purpose. This facilitates greater throughput of communications between partitions, since contention for the system wide lock is avoided.

10                                   b. Type 2 Memory Pages

Allocation of this type of memory page can be requested by a client, for example, to define a buffer for passing message data to a client on another partition. As with Type 1 pages, when a Type 2 memory page is allocated to a given client, a Core Services header is created at the beginning of the page. Figure 32B illustrates the  
15 contents of the Core Services header for Type 2 pages, in accordance with the present embodiment.

The Partition Ownership Mask and Client Group ID fields are identical to the corresponding fields in the header for Type 1 pages. That is, the Partition Ownership Mask indicates which partition(s) have ownership rights to the page, and the Client  
20 Group ID field contains the Client Group ID of the clients that have ownership rights to the page. When the page is first allocated, this field will contain the Client Group ID of the client that requested the allocation.

The DeallocationLock field, like the corresponding field in the header of Type 1 pages, is used to coordinate changes in the ownership of the page. Any partition  
25 intending to effect a change in ownership of a page must first acquire the lock to that page via the DeallocationLock field.

The Type 3 Page Count and Type 3 Page Reference fields relate to an additional feature of the present invention, whereby as part of a request to allocate a Type 2 memory page, zero or more Type 3 pages may be allocated in conjunction with the  
30 Type 2 request in order to satisfy the buffer size in the allocation request. The Type 3 Page Count field specifies the total number of Type 3 memory pages associated with

- 74 -

the Type 2 page, and the Type 3 Page Reference field specifies a location within the Type 2 page that contains references (*i.e.*, pointers) to the associated Type 3 pages.

#### c. Type 3 Memory Pages

5           As mentioned above, this type of memory page is used in conjunction with a Type 2 memory page. A Type 3 page contains client data and is owned by a Client Group; however, the Type 3 page does not contain explicit Client Group information. Rather, the Client Group ownership of a Type 3 page is governed by the ownership of its associated Type 2 memory page, as specified in the Client Group ID field of the  
10   Core Services header of that Type 2 page. The ownership of a Type 3 page is implicitly changed whenever the ownership of its associated Type 2 page is changed.

#### d. Type 4 Memory Pages

          This type of memory page is for static ownership by one or more Partitions.  
15   Unlike Type 1, 2, and 3 memory pages, ownership of Type 4 memory pages is specified in an Allocation Table, described hereinafter. Consequently, all changes to ownership of Type 4 pages require acquisition of the system-wide lock.

### 5. Control Structure Header

20           Figure 20 illustrates the contents of the control structure header 1910, in accordance with the present embodiment. A Version ID field is used to identify the particular release, or version, of the Core Services software running on the computer system. A Shared Memory Status field indicates the status of the shared memory (*e.g.*, "uninitialized," "initializing," "initialized," and "cleanup"). A Partition ID Of Master  
25   Partition field identifies which partition is designated as the "Master" of the shared memory window; the Master partition has added responsibilities for managing the shared memory window, as described more fully below. A Shared Memory Partition Check-In Interval field specifies the time interval at which a partition is required to update certain status information to indicate to other partitions that it is active. A Client  
30   Directory Table Header field contains a pointer to the start of the Client Directory Table

- 75 -

and a lock field that is used to coordinate access to the table in accordance with the lock mechanism of the present invention.

5       The control structure header 1910 ends with information about each of the partitions within the computer system, including the type of operating system executing on the partition (*e.g.*, NT, UnixWare, MCP, etc.) and information needed to issue inter-processor interrupts to the partition.

#### 6. Allocation Structure

10       According to the present embodiment, administration of the shared memory pages is facilitated through an Allocation Table (not shown). Each allocable page in the shared memory window is represented by an entry in the Allocation Table. Each entry indicates whether the corresponding page is "in-use," "available," or references memory that is out-of-use, and may also specify page type. For a Type 4 memory page,  
15       the entry further specifies, in the form of a Partition Ownership Mask like that found within the headers of Type 1 and Type 2 memory pages, which partition(s) have ownership rights in the page. Thus, in this respect, ownership of Type 4 pages is maintained differently than for Type 1, Type 2, and Type 3 pages (where ownership information resides in the Core Services header of the page itself). The Allocation  
20       Table, like the Client Directory Table, itself occupies one or more pages of the shared memory window.

      The Allocation Structure 1912 at the base of the shared memory window controls certain parameters associated with the Allocation Table and other structures. Figure 21 illustrates the contents of the Allocation Structure, in accordance with the  
25       present embodiment. A lock field (Allocation Lock) is used to control access to the Allocation Table. This is the system-wide lock referred to above (as opposed to the individual page locks in the headers of Type 1 and Type 2 pages). Partitions must acquire this lock for any initial allocation of pages. This lock must also be required for any subsequent change in ownership of a Type 4 page, since ownership of Type 4 pages  
30       is maintained in their respective Allocation Table entries. As mentioned above,

- 76 -

however, for subsequent changes in ownership of Type 1 and Type 2 pages, only the individual page locks within the headers of the pages themselves must be acquired. This ability to lock individual pages (Types 1 and 2) facilitates greater throughput between partitions, since contention for the system-wide lock (Allocation Lock) is eliminated.

A Length of Shared Memory Area field specifies the number of allocable pages in the shared memory window. A Shared Memory Page Pointer field provides a pointer to the start of the allocable pages. A Free Page List Header provides a pointer to the start of the Free Page List, and an Allocation Table Header provides the pointer to the start of the Allocation Table.

## 7. Signals

The fundamental unit of communication in this embodiment is a Signal. In the present embodiment, there are two major categories of Signals: (1) inter-Partition Core Services-to-Core Services Signals and (2) inter-Partition Client-to-Client Signals. Core Services-to-Core Services Signals are those that are sent between the Core Services software executing on different partitions. Client-to-Client Signals are those that are sent between clients on different partitions. Each category of Signal has one or more signal sub-types. Each Signal comprises a Core Services Information Section and a Client Information Section. Each of these sections comprises a number of words, the definition of which depends on its type.

For the Core Services-to-Core Services Signal sub-types, the Client Information Section is not defined. All information is contained in the Core Services Information Section. The following Core Services-to-Core Services Signal sub-types are defined in the present embodiment:

(1) Membership Change Signal: whenever a client registers or unregisters with the Core Services software on a partition, the Core Services software must send this Signal to the Core Services software on each other partition that has a client registered to the same Client Group to notify them that its client is registering/unregistering. The Core Services Information Section of the Signal will contain the Client Group ID of the

- 77 -

Client Group to which the client is registering/unregistering with the Group.

(2) Resume Sending Signal: this Signal is used by a receiving partition to alert the Core Services software on a sending partition that it can resume sending Signals to it (the use of this Signal is further described below in conjunction with the description of the overflow flag of each Input Queue).

(3) You Have Been Marked Dead Signal: this Signal is sent by the Core Services software on the Master partition to a partition that the Master has determined is not functioning;

10 With Client-to-Client Signal sub-types, both the Core Services Information Section and the Client Information Section are defined. In the present embodiment, only the following Client-to-Client Signal sub-type has been defined: Signal Delivery Signal. As described in greater detail below, when a client on one partition wishes to send a Signal (and perhaps pass a buffer of message data) to a client on another  
15 partition, the client calls a Send Signal interface of the Core Services API. In response, the Core Services software sends the Signal Delivery Signal to the partition on which the receiving client is running. The Core Services Information Section of the Signal Delivery Signal contains the Client Group ID of the sending and receiving clients and may also contain a handle (*i.e.*, reference) to one or more pages of shared memory that  
20 have been allocated to the client to define, for example, a buffer that contains a shared memory object intended for the receiving partition. Examples of shared memory objects are client messages, client data streams, client events, and Core Services events. The Client Information Section is opaque to the Core Services software, but can be used by the sending and receiving clients for any desired purpose. For example, the  
25 Client Information Section could be used to communicate short messages between clients. In the present embodiment, the Client Information Section comprises a maximum of five (5) words.

## 8. Input Queues and Input Queue Header

30 An input queue mechanism, in combination with the inter-processor interrupt

- 78 -

mechanism described below, is used to signal a recipient partition that data is available.

Each partition has a separate input queue for each other possible partition in the computer system. In the present embodiment, each partition also has an input queue for itself, to be used, for example, in the event that the Core Services software on the partition needs to send a Signal to a client on that same partition. Thus, in the present embodiment, wherein the computer system can be configured into a maximum of eight separate partitions (*i.e.*, each of the eight sub-PODs defining a separate partition), each partition has eight separate input queues (one for each of the other seven partitions and one for itself), for a total of sixty-four (64) input queues. These input queues reside in the portion 1914 of the shared memory control structure 1900, along with a header. Signals will be generated by the Core Services software on one partition and delivered to the Core Services software on another partition via the corresponding input queue between them.

Figure 29 illustrates the contents of the input queue header, in accordance with the present embodiment. An Input Queues Pointer field holds a pointer to the start of the actual input queues. A Number of Input Queues field specifies the number of input queues in the input queue area 1914 (sixty-four in the present embodiment). An Input Queue Length field specifies the length (in words) of each Input Queue. In the present embodiment, the length is specified as 2048 words. An Input Queue Signal Size field specifies the total length of each Signal (Core Services Information Section + Client Information Section). The total size of each Signal is the same and is fixed. Finally, a Number of Signals in Input Queue field specifies the total number of possible Signals that each Input Queue can accommodate at one time.

Figure 30 illustrates the contents of each input queue, in accordance with the present embodiment. As shown, each input queue has a Lock field 3010 which is used by the Core Services software to lock access to the input queue while updating information in the queue, a Count field 3012 that specifies the current number of Signals in the queue, and an Overflow flag 3014 that is used to indicate that the queue has reached capacity but that there are additional Signals to be transferred onto the queue as soon as room becomes available. These fields are followed by space 3016 for a fixed number of Signals (as specified in the Number of Signals in Input Queue field



- 79 -

of the Input Queue Header, *see* Fig. 29).

In the present embodiment, the sixty-four Input Queues are grouped contiguously in the Input Queue area 1914 of the control structure 1900. That is, the first eight Input Queues in the structure belong to the first partition, with successive  
5 groups of eight Input Queues belonging to successive ones of the other seven partitions.

#### a. Preferred Operation

In operation, whenever the Core Services software gets a request from a client to send a Signal to another partition, it builds the Signal based on information supplied  
10 by the client and attempts to place the Signal into an available entry in the appropriate Input Queue for the receiving partition. If no entries are available, then the Overflow flag 3014 of the Input Queue is set to alert the receiving partition that there are Signals waiting to be transferred that could not be transferred because the Input Queue was full, and an error is returned to the client. In such a case, when the receiving partition  
15 subsequently empties the Input Queue, it clears the Overflow flag 3014 and sends a Resume Sending Signal back to the sending partition, to alert the sending partition that it may now transfer any subsequent Signals issued by its clients onto the Input Queue for communication to the receiving partition.

On the receiving side, when the Core Services software on the receiving  
20 partition receives an inter-processor interrupt from a sending partition, it examines the count fields in each of its associated Input Queues to determine which Input Queues have available Signals. When the Core Services software finds an Input Queue with available Signals, it transfers them to a local processing buffer in its exclusive memory window and resets the count in the Input Queue. Each received Signal extracted from a  
25 given Input Queue is then passed to the appropriate client (based on the Client Group ID in the Signal) via a Receive Signal callback interface that all clients are required to implement.

#### b. Alternative Operation

30 In an alternative embodiment, in order to provide more efficient movement of client Signals into the various input queues in response to send requests, the Core

- 80 -

Services software on each partition may set up a partition Send Queue (*i.e.*, buffer) (not shown) in its exclusive memory window for each possible destination partition. In this alternative embodiment, whenever the Core Services software on a partition encounters a full Input Queue that prevents it from placing additional Signals on the Input Queue, it sets the overflow flag in the Input Queue and then queues those Signal requests to the appropriate local Send Queue until entries again become available in the Input Queue.

Additionally, on the receiving side, the Core Services software on each partition may also set up local Client Signal Tank Queues in its exclusive memory window - one for each client that has identified itself to the Core Services software. Each received Signal extracted from a given Input Queue of a receiving partition is transferred into the Client Signal Tank Queue that corresponds to the intended recipient client (again based on the Client Group ID in the Signal). Each Signal in a Tank Queue is eventually passed to the intended recipient client via a call to the client's Receive Signal interface.

The local Send Queues and Tank Queues in this alternate embodiment, in combination with the use of the Overflow flag as described above, are intended to provide efficient and equitable use of the shared memory resources to all of the clients of the Core Services software. Because each client's Signals are queued locally, the Input Queues in the shared memory window are kept open for communication in an efficient manner. No Signals are lost when an Input Queue reaches capacity, and the Input Queues are emptied quickly to minimize the time that Signals wait on a given Send Queue.

### 9. Inter-Processor Interrupt Mechanism

As mentioned above, an inter-processor interrupt mechanism is employed to alert a receiving partition that Signals have been placed in one of its Input Queues by a sending partition. Specifically, in the present embodiment, each partition establishes a single interrupt vector that all other partitions use to send inter-processor interrupts to it. Whenever a sending partition places a Signal in the Input Queue for a given receiving partition that causes the Input Queue to go from an empty state (Count = 0) to a non-empty state (Count > 0), the Core Services software on the sending partition

- 81 -

generates an inter-processor interrupt to one of the processors of the receiving partition.

The processor of the receiving partition responds to the interrupt by calling an interrupt service routine (not shown) of the Core Services software on that partition. Because each partition assigns only a single interrupt vector for receipt of interrupts from the other partitions, the Core Services software on the receiving partition does not know which other partition issued the inter-processor interrupt. Consequently, the Core Services software on the receiving partition must check the Count field 3012 in each of its Input Queues to determine whether any Signals are available in any of those queues.

If an Input Queue has available Signals, the Core Services software transfers those Signals to a local processing buffer in the receiving partition's exclusive memory window and resets the Count field 3012 in the Input Queue. If the Overflow flag 3014 of a particular Input Queue was also set, the Core Services software resets the Overflow flag and sends a Resume Sending Signal back to the sending partition, as explained above. The Core Services software then traverses the local processing buffer, extracting each received Signal, determining the destination client from the Client Group ID in the Signal, and then delivering the Signal to the destination client via the client's Receive Signal callback interface. The Core Services then repeats these steps for each other Input Queue that also has Signals available (*i.e.*, count > 0).

a. Exemplary Intel/Windows NT Implementation

At the processor and operating system levels, inter-processor interrupt mechanisms are both processor and operating system dependent. As one example, the following is a description of how inter-processor interrupts are generated and serviced in accordance with the present embodiment in the case of partitions that employ Intel Pentium-family microprocessors and that execute the Microsoft Windows NT operating system.

In accordance with the present embodiment, the Hardware Abstraction Layer (HAL) of the Microsoft Windows NT operating system is modified so that during initialization of the HAL on a given partition, the HAL will first select an inter-processor interrupt vector for receipt of shared memory inter-processor interrupts by that partition. An interrupt vector is a number that is assigned to an incoming interrupt

- 82 -

hardware signal by the HAL of the Windows NT operating system. For example, interrupt vectors are typically assigned by the HAL to the various device I/O hardware interrupt signals on a system. An inter-processor interrupt is a specialized type of hardware interrupt signal that is sent from one processor to another (as opposed to from an I/O device to a processor). As with general I/O interrupts, the HAL must also assign vectors to any inter-processor interrupt signals (from the same number space that the I/O interrupt vectors are chosen). Thus, in the present embodiment, the modified HAL assigns an interrupt vector for the inter-processor interrupts that will be received by the local Core Services software on that partition to alert the software that one or more  
5 Signals are available in at least one of its Input Queues.  
10

In the case of an Intel microprocessor, inter-processor interrupts are actually generated and received by an advanced programmed interrupt controller (APIC) associated with the processor. The APIC associated with the sending processor generates a hardware signal to the APIC associated with the receiving processor. If  
15 more than one processor is to receive the interrupt, then the APIC of the sending processor will generate a hardware signal to the APIC of each intended recipient. The APIC of each receiving processor receives the hardware signal and delivers the corresponding interrupt vector to the processor for handling.

Further according to the present embodiment, in addition to assigning an  
20 interrupt vector for the receipt of inter-processor interrupts from other partitions, the modified HAL will also designate one or more processors in its partition to handle such interrupts. In the present embodiment, in the case of a partition that comprises more than one sub-POD, the designated processors must be members of a single one of those sub-PODs (this is a limitation imposed by the present embodiment of the computer  
25 system platform and may not be a limitation in other embodiments). When more than one processor on a sub-POD has been designated, an incoming interrupt will be received in the local APICs of each of those processors. The APICs will then arbitrate to determine which one of the processors will handle the interrupt. Further details concerning this arbitration process are provided in the Pentium Pro Family Developer's  
30 Guide: Volume 3, available from Intel Corporation. Additional information concerning APICs can be found in the Intel MultiProcessor Specification, version 1.4, also

- 83 -

available from Intel.

Still further according to the present embodiment, when the Core Services software is initialized on a partition, the Core Services software queries the HAL of the NT operating system on that partition through a custom interface to obtain the interrupt  
5 vector and the information concerning the processors designated by the HAL to handle shared memory inter-processor interrupts incoming to that partition. The Core Services software then stores this information in the Partition Information section of the Control Structure Header 1910 (see Fig. 20). This makes the information accessible to the Core Services software on other partitions. The Core Services software will then supply the  
10 HAL, through another interface, a reference to an interrupt service routine that is part of the Core Services software. If a designated processor on that partition receives an inter-processor interrupt with the designated interrupt vector, it will execute the interrupt service routine, allowing the Core Services software to respond to the interrupt.

In operation, in order to generate an inter-processor interrupt to notify a  
15 receiving partition that a Signal has been placed in one of its Input Queues, the Core Services software on the sending partition looks up the inter-processor interrupt information of the intended recipient partition in the Control Structure Header 1910. The Core Services software then calls another custom interface to the HAL on its partition, supplying the HAL with the inter-processor interrupt information for the  
20 receiving partition. With this information, the HAL on the sending partition manipulates the registers on the APIC of one of its processors to cause an inter-processor interrupt signal to be generated from its APIC to the APICs of each processor designated by the HAL on the receiving partition to receive such inter-processor  
interrupts. Those APICs on the receiving partition will then arbitrate to handle the  
25 interrupt, and the processor that wins the arbitration will invoke the interrupt service routine of the Core Services software on the receiving partition.

#### b. Alternative Embodiment - Multiple Interrupt Vectors

In the embodiment described above, each partition is assigned a single interrupt  
30 vector for receipt of shared memory inter-processor interrupts from any of the other partitions. Because of this, a receiving partition does not know which other partition

- 84 -

generated the received interrupt. Consequently, the receiving partition must examine each of its Input Queues, in turn, to ensure that it receives the Signal(s) from the sending partition that generated the interrupt.

As an alternative embodiment, each partition may assign a separate interrupt vector for receipt of shared memory inter-processor interrupts from each other partition.

A sending partition would then generate an inter-processor interrupt to a receiving partition using the corresponding interrupt vector assigned to it by the receiving partition. An advantage of this embodiment is that a receiving partition would know from the interrupt vector which other partition generated the incoming interrupt. The Core Services software on the receiving partition could then access the appropriate Input Queue to retrieve the incoming Signal(s), without having to cycle through all of the Input Queues as in the embodiment described above.

#### 10. The Core Services API

In order to provide the functionality described above to clients of the Core Services software, the Core Services software has a defined application programming interface (API) that provides interfaces (*i.e.*, callable methods) that a client can call to invoke the services of the Core Services software. The following is a list of interfaces provided as part of the Core Services API to perform the functions described above:

Initialize Client Software - this interface is used by a client to identify itself to the Core Services software. The Core Services software returns a Client Reference identifier to the Client.

Uninitialize Client Software - this interface is used by a client to inform the Core Services software that it will no longer participate as a user of shared memory.

Register Client - this interface is used by a client to register with the Core Services software as a member of a given Client Group. Each client must register before it is allowed to request that any shared memory be allocated to it. The client supplies the desired Client Group Name and its Client Reference identifier as part of the call. The Core Services software will then make the appropriate changes to the Client Directory Table to reflect the addition of this client to the desired Client Group.

- 85 -

The interface then returns the Client Group ID to the client.

Unregister Client - this interface is used by a client to unregister from a particular Client Group.

Allocate Shared Memory - this interface is used by a client to request allocation  
5 of one or more pages of the shared memory window. The client supplies its Client  
Group ID and the buffer size (in bytes) that it is requesting. The Core Services  
software locks the Allocation Table, determines whether enough pages to satisfy the  
request are available in the Free Page List, and then removes those pages from the Free  
Page List. The Allocation Table entries for each allocated page are updated to reflect  
10 that the pages are "in use." For Type 1 and Type 2 pages, a Core Services header is  
created in the page which, as explained above, indicates ownership of the page by  
partition and client. Any Type 3 pages associated with a Type 2 page are referenced in  
the header of the Type 2 page. For Type 4 pages, partition ownership is reflected in the  
corresponding Allocation Table entries. The Core Services software then returns a  
15 handle to the client that the client subsequently uses to reference the pages that  
comprise the allocated buffer.

Deallocate Shared Memory - this interface is used by a client to request that all  
pages associated with a given handle be deallocated. If the requesting partition is the  
only owner of the pages to be deallocated, then the pages are returned to the Free Page  
20 List (the system-wide lock must be acquired in order to do this). If not, then only the  
ownership information (in the Core Services header of Type 1 and Type 2 pages, or in  
the Allocation Table entries for Type 4 pages) is updated .

Send Signal - this is the interface that clients use to have a Signal inserted into  
the Input Queue of a receiving partition. The client calling this interface provides (i)  
25 the Client Group ID of the Client Group of which it and the receiving client(s) are  
members, (ii) an indication of which partitions have a client that will receive the Signal  
( because only one client on a given partition can be a member of a particular Client  
Group, this indication and the Client Group ID are the only pieces of information  
needed to identify the receiving client on each partition), (iii) the actual information to  
30 be supplied with the Signal in the Client Information Section, (iv) a flag indicating  
whether this is a point-to-point or multicast Signal (point-to-point has only one

- 86 -

receiving partition, whereas multicast has multiple receiving partitions), and (v) an optional handle to a shared memory object, such as, a buffer (one or more shared memory pages) containing a client message. In response to a Send Signal call, the Core Services software will (i) build the Core Services Information and Client Information Sections of the Signal, (ii) check the status of shared memory, (iii) insert the Signal in the appropriate Input Queue, and if the Signal was placed in an empty Input Queue, (iv) generate an inter-processor interrupt on the receiving partition. If an Input Queue of an intended recipient partition is full, or the intended recipient partition is down, appropriate error indications will be returned.

10

### 11. Interfaces Supplied by Clients

In addition to the foregoing interfaces supplied by the Core Services software, any client of the Core Services software must implement certain callback interfaces that the Core Services software can invoke to notify the clients of certain events. In the present embodiment, these callback interfaces include interfaces for (i) notifying the client that a Signal has been received ("the Receive Signal interface"); (ii) notifying the client that there has been a membership change in its Client Group; (iii) notifying the client that shared memory is "up" or "down," (iv) notifying the client that the Core Services software is shutting down, and (v) notifying the client that one or more shared memory pages has a memory error.

15

20

### 12. Exemplary Operation

To further illustrate the operation of the interrupt-driven shared memory mechanism described above, Figures 31A and 31B comprise a flow chart that illustrates the steps performed by the clients and Core Services software on two partitions in order to communicate a message from one client to the other.

25

30

Figure 31A illustrates the steps that are performed on the sending partition. At step 3110, the client calls the Allocate Shared Memory interface of the Core Services API, requesting a buffer that will be used to transfer the message to the client on the receiving partition. In this example, the client requests that a Type 2 page be allocated.



- 87 -

The client provides the required buffer size with the request. In response, at step 3112, the Core Services software determines the number of shared memory pages that will be required to satisfy the buffer request (*i.e.*, whether any additional Type 3 pages will be allocated with the Type 2 page). At step 3114, the Core Services software (i) acquires the system wide Allocation Lock, (ii) determines from the Free Page List whether the required number of pages are available and, assuming that they are, (iii) allocates the pages to the client. The Core Services software updates the Allocation Table to indicate that the pages are "in use," and then indicates ownership of the pages in the Core Services header of the Type 2 page. At step 3116, the Core Services software returns a handle to the allocated pages to the client and releases the Allocation Lock.

Next, at step 3118, the client fills the allocated buffer with the message data. Then, at step 3120, the client calls the Send Signal interface of the Core Services API, providing (i) the Client Group ID and receiving partition (which together identify the receiving client), (ii) any information to be provided in the Client Information Section of the Signal, (iii) the handle to the allocated buffer, and (iv) a flag indicating that this is a point-to-point request, as opposed to a multicast request. Recall from above that the client has the option to send a Signal to multiple partitions using the multicast feature of the present invention.

In response to the Send Signal request, at step 3122, the Core Services software identifies the appropriate Input Queue based on the designated receiving partition. The Core Services software then locks the Input Queue (step 3124), increments the Count field (step 3126), and builds the Signal in the Input Queue (step 3128) as an entry in that queue. Next, if the Input Queue was previously empty (*i.e.*, the Count has gone from zero to one) (step 3130), then the Core Services software generates an inter-processor interrupt on the receiving partition (step 3123). If the Count field of the Input Queue was already non-zero, the Core Services software does not need to generate an interrupt. The Core Services software then releases the lock on the Input Queue (step 3131 or step 3133).

Referring now to Figure 31B, the steps performed on the receiving partition are shown. At step 3134, one of the APICs on the pre-designated sub-POD of that partition arbitrates for, and delivers to its processor, the inter-processor interrupt generated by

- 88 -

the sending partition. In response, the processor calls an interrupt service routine (not shown) of the Core Services software. As part of the interrupt service routine, the Core Services software begins examining, at step 3136, the first of its Input Queues (in the present embodiment, there are eight Input Queues for each partition). At step 3138, the  
5 Core Services software examines the Count field of the Input Queue. If the Count is zero, then no Signals have been sent from the sending partition that corresponds to that Input Queue, and the Core Services software proceeds to the next Input Queue.

If, however, the Count of a given Input Queue is greater than zero, then Signals are present and control passes to step 3140. At step 3140, the Core Services software  
10 copies each Signal in the Input Queue to a local processing buffer, and then at step 3142, resets the Count to zero. Next, at step 3143, the Core Services software determines whether the Overflow flag in the Input Queue is set. If the Overflow flag is set, the Core Services software resets the Overflow flag and then sends a Resume Sending Signal to the sending partition, thus alerting the sending partition that the Input  
15 Queue is no longer full.

Next, steps 3144 and 3146 are performed for each Signal copied into the local processing buffer. Specifically, at step 3144, the Core Services software extracts a Signal from the local processing buffer. At step 3146, the Core Services software calls the Receive Signal interface of the recipient client (as identified by the Client Group ID  
20 in the Signal), passing the Client Information Section and the handle to the allocated buffer associated with the Signal (if there is one). At step 3148, the client processes the Signal, including, for example, using the handle to access message data in the referenced buffer. Steps 3144 and 3146 are repeated for each Signal in the local processing buffer. When this is done, the Core Services software repeats steps 3136  
25 through 3146 for each of its other Input Queues. Although not illustrated in Figure 31B, in the present embodiment, the Core Services software on the receiving partition continues to cycle through its Input Queues until it has made a complete pass through all of the Input Queues without finding any waiting Signals (*i.e.*, none with a count > 0). Input Queue processing then stops until another inter-processor interrupt is  
30 received.

An additional aspect (not shown) of the sending and receiving processes is the

- 89 -

deallocation of the allocated shared memory pages. When a sending client that has requested allocation of a buffer (*i.e.*, one or more shared memory pages) transfers the buffer to a receiving partition by passing its handle to the receiving partition via a Signal, the sending partition has the option of either (i) extending ownership rights to the pages of the buffer to the receiving client (in which case both clients will have ownership rights), or (ii) transferring ownership rights to the receiving partition (in which case the sending client relinquishes ownership). Regardless of which option is chosen, at some point, a client may wish to deallocate the allocated pages. This is done using the Deallocate Shared Memory interface. Specifically, a client calls the Deallocate Shared Memory interface, passing the handle to the pages to be deallocated. If no other clients are owners of those pages, then the pages are returned to the Free Page List and their corresponding Allocation Table entries are updated to reflect their availability. If, however, other clients also have ownership rights to those pages, then the pages cannot yet be returned to the Free Page List. Rather, the Core Services software locks down the pages and updates the ownership information in the Core Services header of the Type 2 page.

### 13. Other Functions

In addition to the foregoing, the following additional functions of the interrupt-driven shared memory management mechanism are provided:

#### a. Initialization and Shut Down

When Core Services software begins execution on a partition, it first confirms the availability and status of the shared memory window, and then invokes appropriate platform interfaces to get the following information: the physical address and size of the shared memory window, the partition identifier (each partition has an associated identifier), the information needed by other partitions to generate inter-processor interrupts to this partition, and the host operating system type and version running on the partition. The Core Services software stores a copy of this information in the exclusive memory window of its partition and in the various fields of the shared

- 90 -

memory control structure 1900, such as, for example, the Partition Information field of the control structure header 1910 and the Length of Shared Memory Area field of the Allocation Structure 1912.

5           In order for a partition to join other partitions in accessing and using the shared memory window, the partition must make itself known to the other partitions using the shared memory window. If there is no current Master partition, then they must arbitrate among themselves to elect a Master partition. For this purpose, Core Services has a 'Check In' mechanism. The 'Check In' mechanism enables each partition to determine  
10          the validity of the Shared Memory Status field in the Control Structure Header without using a lock, and to dynamically elect a new Master when there is no active Master.

          It is also the responsibility of the Core Services software to exit the shared memory window cleanly whenever a partition voluntarily leaves the shared memory  
15          window. This is true for both the Master partition and the non-Master partitions. The common responsibilities of any departing partition are: (i) to notify its local clients that the shared memory window is going away by calling the appropriate client callback interface, (ii) to unlock any data structures that it has locked (*e.g.*, Allocation Table, Input Queue, etc.), (iii) to clean up its Input Queues, (iv) to deallocate any shared  
20          memory pages that it owns, (v) to return any local memory that it owns, and (vi) to change its status in the Control Structure Header 1910 to "Uninitialized".

          If the departing Partition is the Master partition and there are no other alive partitions, then it shuts down the shared memory window with a notification sent to the  
25          MIP. If the departing partition is the Master partition and there is at least one other partition still communicating with the shared memory window, then a new Master partition is chosen by the remaining active partitions.

- 91 -

b. Master Partition Duties

The Master Partition has specific responsibilities when shared memory is initialized, when a non-Master partition dies, and when shared memory shuts down. The following duties are reserved for the Master Partition:

- 5 (1) initialize shared memory structures, including the Control Structure Header, the Allocation Structure, the Allocation Table, the Free Page List, the Input Queue Header, the Input Queues, the Client Directory Table Header, and the Client Directory Table;
- (2) perform house cleaning operations on shared memory structures and in-use  
10 shared memory pages when a partitions dies; and
- (3) perform house cleaning operations on shared memory structures when shared memory shuts down.

c. Duties of Non-Master Partitions

15 All the partitions, including the Master partition, have the following duties:

- (1) monitor the status of the other partitions at the predefined Shared Memory Partition Check In Interval;
- (2) determine if a new Master partition needs to be chosen;
- (3) update the appropriate areas in the shared memory structures and deallocate  
20 any shared memory pages it owns if it chooses to leave the shared memory window; and,
- (4) deallocates any shared memory pages owned by a client, if the client withdraws its participation in the shared memory window or the client fails.

25 As described herein, the program code that implements the interrupt-driven shared memory communication mechanism of this alternative embodiment is implemented as a combination of both operating system code (*e.g.*, the modification to the HAL) and a separate computer program (*e.g.*, the Core Services software). It is understood, however, that in other embodiments, the program code could be  
30 implemented either entirely as operating system code or entirely as a separate computer program without deviating from the spirit and scope of the present invention as defined

- 92 -

by the appended claims. Moreover, the program code can also be implemented in hard-wired circuitry or a combination of both hard-wired circuitry and software code. As mentioned above, the term "program code" is intended to encompass all such possibilities.

5

#### **IV. Exemplary Uses of the Computer System and Methods of the Present Invention to Facilitate Communications Between Partitions**

10 Exemplary uses of the computer system described above, including its shared memory management features, to facilitate communication between operating systems and/or applications running under the operating systems, are described below. Exemplary embodiments of these uses are described below for purposes of illustration, and not limitation. Alternate embodiments (including equivalents, extensions, variations,  
15 deviations, et cetera, of the embodiments described herein) will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein. The invention is intended and adapted to include such alternate embodiments.

##### **A. A Shared Memory Device Driver**

A shared memory network driver interface specification (NDIS) device driver, as described below, can be implemented to allow standard off-the-shelf applications to operate on the multi-partition system described above. The shared memory NDIS device  
25 driver provides standard networking and/or clustering interfaces with faster bandpass and with lower latency than on an analogous LAN configuration, for example. This shared memory NDIS device driver is built upon, and takes advantage of, the Core Services software of the interrupt-driven shared memory management mechanism described above in Section III.B.

30

- 93 -

Figure 18 illustrates the exemplary shared memory NDIS device driver 1802. The unshaded boxes represent standard Windows NT components. The shaded boxes represent components that can be implemented as part of the invention.

5       The shared memory NDIS device driver 1802 supports an upper-edge interface and a lower-edge interface. On the upper-edge interface, the shared memory NDIS device driver 1802 supports standard NDIS interfacing to standard network protocol drivers. The shared memory NDIS device driver 1802 functions as an NDIS layered driver. More specifically the shared memory NDIS device driver 1802 conforms to NDIS Miniport  
10 interfaces and supports any network protocol using the NDIS interfaces to communicate over NDIS device drivers. For example, TCP/IP and SPX/IPX protocols can be implemented.

      The lower-edge interface for the shared memory NDIS device driver 1802 is a  
15 private interface to the Core Services software described in Section III.B., which directly supports the global shared memory capabilities. The interface includes a hybrid of normal layered IO driver interfaces (IRPs) and tightly coupled IO driver interfaces (direct procedure call). The IRPs are used for asynchronous functions. The tightly coupled IO driver interfaces are used for synchronous functions.

20       The main function of the shared memory NDIS device driver 1802 is to map the NDIS interface onto the Core Services API. Local system buffers containing networking packets (NDIS packets) are passed through the NDIS interface to the shared memory NDIS device driver 1802. The shared memory NDIS device driver 1802 copies the network  
25 packet from the local system buffer (in a partitions exclusive memory window) into a shared memory buffer. A reference to the shared memory buffer is queued to the appropriate shared memory NDIS device driver in another partition as selected by the destination MAC address in the network packet. Packets with a broadcast or multicast MAC address are copied into as many shared memory buffers as necessary to send directly  
30 to each partition supporting a device driver in shared memory NDIS device driver 1802's shared memory group, thus simulating a broadcast/multicast. Buffers received from shared

- 94 -

memory will be repackaged into NDIS packets and presented to the NDIS interface where they are processed by network protocol drivers. The NDIS packets are returned to shared memory NDIS device driver 1802.

5           The shared memory NDIS device driver 1802 maintains a list of shared memory buffers for each partition, called a SendList, to reduce the overhead of allocating and deallocating shared memory buffers via the Core Services software. Shared memory buffers are selected from the SendList for sending network packet information to another partition. The receiving partition will have a RcvList of handles corresponding to the  
10           originating partitions SendList. When the receiving partition is finished with the message processing, it sends a message indicating that the buffer should be returned to the available state in the SendList. When the number of buffers in the SendList drops below a minimum value, additional buffers are allocated from the Core Services software. When the number of buffers in the SendList is at a maximum, and not all in use, buffers are deallocated back  
15           to the Core Services software. The minimum and maximum SendList sizes have pre-determined default values in the code, but can be overridden by setting specific keys in a registry.

          The shared memory NDIS device driver 1802 uses the Core Services software on  
20           its partition 1804 to simulate a FDDI LAN between all the partitions that are running copies of the shared memory NDIS device driver 1802. The shared memory NDIS device driver 1802 supports the basic semantics of an FDDI LAN. This includes point to point messaging, broadcast messaging, multi-cast messaging and 4491 byte message sizes.

## 25           ***B. Maintaining an Appearance of Communications by Wire***

          In another exemplary application of the computer system and its global shared memory management, sharing of memory between partitions (*i.e.*, Pods, sub-Pods or operating systems) is achieved while maintaining an appearance of communications by  
30           wire. This permits conventional applications programs, conventional application program



- 95 -

interfaces (APIs), and conventional communications hardware and software to be used to send data to shared memory. This application is built upon the mechanism described above in Section III.A., in which inter-partition communications are managed in accordance with a polling technique.

5

Figure 22 is an exemplary configuration of the computer system 200 of the present invention, including additional software components needed to achieve the appearance of communications by wire between partitions or operating systems. In figure 22, two partitions 2202a and 2202n are shown, each of which may, for example, include a single sub-Pod. Each sub-Pod 2202 operates under control of a separate operating system 2206. Operating systems 2206 can be separate instances of the same operating system or they can be different operating systems. One or more application programs 2208 can run on each partition 2202 under the operating system 2206 that operates on that partition.

15

One or more application program interface (API) modules 2210 can be associated with one or more application programs 2208, for sending messages. For example, on sub-Pod 2202a, application program 2208a can initiate a message send operation using API 2208a. API 2208a prepares the message for input to a network communications interface module 2212.

20

Network interface communications interface module 2212 can be a conventional system that interfaces partitions with one another, such as through a network. Network interface communications module 2212 formats messages for transmittal to other partitions 2202 through a network driver 2216 and over a conventional network-type wire 2214. In an exemplary embodiment, network interface communications module 2212 outputs messages on lines 2220a and 2220b as if they were destined for a conventional network-type wire transmittal system 2214. Thus, up to this point, sending of messages from partitions 2202a is performed in a conventional manner.

30

Instead of sending all messages on lines 2220a and 2220b from network communications interface module 2212 to a conventional network driver 2216, messages

- 96 -

intended for shared memory 160 are handled through a shared memory driver 2218. In an exemplary embodiment, a destination address is associated with each message. If an address corresponds to a computer or other destination that is coupled to wire 2214, then the message is sent to wire 2214 through network driver 2216. If, however, the address  
5 corresponds to an address in shared memory 160, the message is directed to shared memory driver 2218.

Shared memory driver 2218 receives and reformats messages for transmittal to, and storage in, shared memory 160. Reformatting can include, for example, reformatting  
10 messages into a standard format that can be recognized by application programs 2208 running on other partitions 2202. Reformatting can also include, for example, reformatting in accordance with specifications associated with shared memory 160.

Referring to Figure 23, further details of system 2200 are illustrated. In this  
15 exemplary embodiment, operating system 2206a on partition 2202a is illustrated as a 2200 operating system, commercially available from Unisys Corporation, and operating system 2206n on partition 2202n is illustrated as a Windows NT or a UNIX operating system.

In the exemplary embodiment of Figure 23, network communication interface  
20 modules 2212 include one or more software modules 2310 that implement a conventional transport layer (i.e., layer 4) of an Open Systems Interconnection (OSI) seven-layer communications model. The OSI seven-layer communications model is well known to persons skilled in the relevant art(s). The transport layer can be implemented using a number of different protocols, including a Transmission Control Protocol (TCP), and a  
25 User Datagram Protocol (UDP). The selected protocol will determine the reliability of, and the potential for duplication during, the subsequent communication operation. In an exemplary embodiment, TCP can be utilized to provide reliable unduplicated data delivery.

The software module that implements the transport layer 2310, interfaces with a  
30 software module that implements a network layer 2312, which is layer 3 of the seven-layer OSI protocol. This can be performed using the industry-recognized Internet Protocol (IP)

- 97 -

and Internet Control Message Protocol (ICMP), for example. IP dictates the protocol used for data transmission. ICMP defines the manner in which error handling and analysis is performed.

5           The software module(s) that implements layer 3 2312 interfaces with a communications handler 2314. Communications handler 2314 formats message data into packets. A format can comply with a selected one of a number of communications protocols. These protocols can include, for example, Ethernet, Token Ring, Fiber Distributed Data Interface (FDDI), Asynchronous Transfer Mode (ATM), etc. In an  
10       exemplary embodiment, an Ethernet Handler, which implements an Ethernet protocol, is used.

          After a message is formatted within local memory, communications handler 2314 calls a device driver. During a "normal" communication scenario, an I/O Driver is called  
15       to perform communications via a network. In an exemplary embodiment, this is a network input/output device driver (NIOP) 2316, commercially available from Unisys Corporation. NIOP 2316 implements layers 2 and 1 of the OSI model, which are the data link and physical layers of the model, respectively.

20           When communication is to be performed via shared memory 160 instead of over a network, a shared memory driver 2218 is called. For example, on partition 2202a, when communication is to be performed via shared memory 160 instead of over a network, communications handler 2314 can call a HMP Shared Memory Driver 2318 instead of NIOP Driver 2316. Communications handler 2314 does not need to distinguish between a  
25       call to NIOP Driver 2316 and a call to HMP Shared Memory Driver 2318. From communications handler 2314's point of view, all messages are transferred over a network. The operating system decides which of the two types of calls is to be made, as will be discussed further below. The functionality included within the HMP Shared Memory Driver is described below.

- 98 -

The functionality included in the 2200 operating system software modules on partition 2202a is included in similar modules residing in the NT or Unix operating system of partition 2202n. In Figure 23, these modules can include an API 2210n (shown as Winsock/Sockets), and network communications interface modules 2212 (shown as TCP/UDP/IPS 2310n, IP/ICMP 2312n, and Ethernet Handler 2314n). Communications with memory 160 is through HMP Shared Memory NIC device driver 2320. As with the 2200 operating system software modules, the layers of the software that interface to the Applications Program, including the API and the communications software, do not distinguish between network or shared-memory communications. These software components view all communications operations as occurring over a network.

Figure 24 provides further details of the HMP Shared Memory Driver 2320 as implemented in a Windows NT environment in accordance with an exemplary embodiment of the invention. In Figure 24, a NT User Application 2410 interfaces to a dynamic link library 2412. Dynamic link library 2412 interfaces with a Windows Socket 2414. Windows Socket 2414 interfaces with a Transport Driver Interface (TDI) 2416, which is a Microsoft-defined API for NT systems. API 2416 interfaces to a TCP/IP module 2418 which performs layers three and four of the OSI communications model. TCP/IP module 2418 can interface with a device driver via an API 2420 designed according to a Network Driver Interface Specification (NDIS) developed by the Microsoft and 3Com Corporations. The device driver can be, for example, an off-the-shelf driver, such as a COTS Ethernet Device Driver 2422, which performs message transmission over an Ethernet network, or may be HMP Shared Memory NIC Device Driver 2320. When the API 2420 makes a call to a device driver, API 2420 does not distinguish between the two types of calls, and all communications appear to be performed via a network.

HMP shared memory NIC device driver 2320 can include, for example, VLAN 2424, CONTROL 2426, SHM 2428, and BIOS 2430 modules. Operation and functionality of these modules is described below.

- 99 -

Figure 25 is a process flowchart illustrating further details of the operation of the software component illustrated in figures 22-24, in accordance with the present invention. The process begins at step 2510 where an application program builds a message and associated header information in local memory.

5

In step 2511, the application program calls an associated API. The program passes the API the length of the message, the IP address of the target host, and one or more pointers to the message data. If the message is to be passed over a network, the IP address specifies a device driver such as the NIOP (on the 2200 operating system side) or an Ethernet LAN NIC Device Driver (on the NT or UNIX side). If the message is to be passed via shared memory, the IP address indicates that an associated HMP Shared memory driver is to be used.

10

In step 2512, software modules which perform layers 3 and 4 of the OSI model add various headers to the message and format the message data to conform with the requirements of the selected communications protocol. For example, the Ethernet protocol requires that a single message transmission may contain no more than 1500 bytes. A longer message must therefore be formatted into multiple buffers to be sent via multiple message transmissions.

15

20

In step 2514, a communications handler (which, in an exemplary embodiment, is an Ethernet handler) makes a call to the Operating System (OS) for the address of the device driver. One skilled in the relevant art(s) will recognize that other protocols could be employed, including, for example, protocols with a larger network data packet size.

25

Generally, the communications handler will connect to a device driver before any application messages are received for transmission. The communications handler will send its own 'broadcast' message out over the network asking everyone to respond with their identity, which for TCP/IP, results in IP addresses being returned. This is how the communications handler knows what IP addresses can be accessed.

30

- 100 -

In step 2516, the operating system selects a device driver address associated with the specified IP address, and passes the address to the communications handler. In an exemplary embodiment, the operating system maintains a table which maps IP addresses to various device drivers. The device driver address may specify a device driver which performs network communications (such as the NIOP or the Ethernet LAN NIC Drivers). Alternatively, the device driver may specify a device driver which performs communications via shared memory. The communications handler is not able to distinguish between the two types of addresses. The 2200 operating system device driver for shared memory can be adapted from a 2200 operating system NIOP, as described in U.S. Patent Number 5,659,794, assigned to Unisys.

In steps 2518-2528, when the address indicates communication is to be performed via shared memory, an HMP Shared Memory Driver (2200 operating system) 2318 or an HMP Shared Memory NIC Device Driver (NT/UNIX) 2320 is called. The called driver first maps the target host ID to one of the nodes. This determines which one of queues within the sending nodes' Output Queue will be utilized.

In step 2518, the called driver determines whether the queue for the target (receiving) system requires resetting. If the queue for the target system requires resetting, processing proceeds to step 2526 where the sending system (or sending "node") discards the message, and sets a Need\_Reset flag in the queue for the target system (or target "node"). When the Need\_Reset flag is set, a reset procedure can be performed.

Where a TCP protocol is used instead of UDP, the message can be discarded without resulting in message loss. This is because TCP waits for an acknowledge from the receiving system indicating that the message has been received. This is tracked using message IDs. Each message is retained in the local storage of the sending system until an associated acknowledge is received. If an acknowledge is not received within a predetermined period of time, another call is made to the operating system to re-send the message. If UDP is utilized instead of TCP, the message would be lost since UDP does not track the receipt of acknowledges from the receiving system.

- 101 -

Typically, the sending application decides whether UDP or TCP is used. This decision is transparent to shared memory. In an exemplary embodiment, the shared memory of the present invention supports UDP, TCP and higher layer protocols that connect with the device driver that handles shared memory. From a communications handler point of view, shared memory of the present invention is just another LAN that does not have very many nodes connected.

If the target queue does not require resetting, processing proceeds to step 2520, where the sending system checks to determine if the target queue is full. In an exemplary embodiment, this is done by comparing the value stored in the appropriate Enqueued\_offset (in the sending node's output queue) to the associated Dequeued\_offset (in the receiving node's input queue). If putting a new entry in the target output queue will cause the Enqueued\_offset to be equal to the Dequeued\_offset, then the target output queue is full.

When the target output queue is full, processing proceeds to step 2528 where the message is discarded. The message can be re-sent later, as discussed above with regard to steps 2518 and 2526.

When the target output queue is not full, processing proceeds to step 2522 where a message buffer in shared memory is obtained from the sending node's message buffer pool. One skilled in the relevant art(s) will recognize that this can be implemented in a variety of ways. In an exemplary embodiment, a memory management module is associated with the Shared Memory Device Driver on each node to keep track of empty buffers.

Preferably, for each Output Queue, a buffer pool including, for example, at least, 511 buffers, will be available. Each buffer can be, for example, 427 8-byte words in length. In an exemplary embodiment, each buffer pool starts on a 4K word page boundary, wherein each word is 8 bytes long. That is, a new buffer pool may start on every eighth 4K-byte page boundary. This allows for more efficient memory management.

- 102 -

For example, each buffer pool can be  $511 * 427 * 8 // 4096 = 1,748,992$  words long, where 511 is the number of queue entries, 427 is the number of words needed to handle a 1500 byte long message and an extra header needed to handle the 2200 operating system requirements. 1500 divided by four equals 375 plus 50 maximum parts and two for buffer and header length for a total of 427. Eight is for the maximum number of partitions and 4096 is to round it up to a page boundary for protection reasons.

After a buffer is obtained, processing proceeds to step 2524, where the message is placed in the output queue by copying from local memory to the shared memory buffer. During this process, a header is generated which serves as the header defined in physical layer, layer 1, of the OSI model.

The header in the shared memory buffer can be viewed as a physical layer because the MAC and LLC layers will be on the message when received by the shared memory device driver. These headers will remain because at least the LLC layer is needed for potential routing at the receiving node. The header in the buffer is necessary because of the different memory access characteristics of the 2200 style processor and the Intel platforms and represents how the data is at the physical layer.

When a 2200 operating system is performing the message send operation, the Block Transfer Pack (BTP) hardware instruction is used to move the message data from local to shared memory. This instruction converts the message data from 9-bit bytes to 8-bit bytes, performs a zero-fill operation, and big endian (2200 style processor) to little endian (Intel) conversion. Alternatively, this conversion could be performed in software.

In an exemplary embodiment, the message is added to the Output Queue by adding the pointer to the message buffer in the appropriate location within the Output Queue, then incrementing the appropriate Enqueued\_offset with the sending node's Output Queue. The pointer is an offset from the start of the sending node's buffer area. Preferably, offsets are used instead of real or virtual addresses so that all nodes are able to get to the same address



- 103 -

in the shared memory. (A receiving node's virtual or real addresses are not necessarily mapped to the same location in memory as another node's virtual or real addresses.)

As previously described with regard to Figures 23 and 24, when a 2200 operating  
5 system node is sending a message, a call is made to the operating system for a device  
driver address. The 2200 operating system uses the IP Address to decide whether a NIOP  
device driver or HMP Shared Memory Driver should be utilized during the  
communications operation. If an NT node is sending a message, similar functionality is  
provided. The VLAN component receives the message-send call from NDIS. VLAN  
10 passes this call to CONTROL, which determines whether the IP address associated with  
the message-send operation is mapped to the Ethernet Device Driver, or to the SHM  
Device Driver, and makes the appropriate device call. The SHM module performs the  
functionality illustrated in steps 2518-2528.

15 In order to receive a message, each node in the system performs a loop that checks  
the Output Queues for each node in the system. In an exemplary embodiment, each node  
performs this check as if the system is fully configured with the maximum number of eight  
nodes, even if fewer nodes are available. The Output Queues of the nodes which are not  
available can be initialized so that it appears that no messages are available. Each node  
20 checks its own Output Queue to determine if it is sending a message to itself, even though  
this will generally not occur. These are design decisions that can be implemented to  
simplify the code.

Alternatively, the number and identity of the available nodes can be communicated  
25 to each node during system initialization so that only the output queues of nodes that are  
actually present are checked. In this embodiment, each change in the number of nodes  
participating in shared memory is communicated to the participating nodes when the  
change occurs.

30 Figure 26 illustrates an exemplary message receiving process performed for each  
partition. The process beings at step 2610, where a message receiving node checks a

- 104 -

Need\_Reset flag in another sub-Pod's output queue. For example, Node 0 checks the Need\_Reset flag in the Node-1-to-Node-0 Queue in the Node 1 Output Queue. If the Need\_Reset flag is set, processing proceeds to step 2612, where an initialization sequence is performed.

5

If the Need\_Reset flag is not set, processing proceeds to step 2614 where the message receiving sub-Pod compares an appropriate Enqueued\_offset flag with one of its own Dequeued\_offset flags in its own Output Queue. For example, in Figures 16A and 16B, Node 0 compares the Enqueued\_offset flag in the Node-1-to-Node-0 Queue in the Node 1 Output Queue to the Dequeued\_offset for Node 1 in its own Output Queue (in Word 1 of the Dequeued\_offsets). If the values stored within the two fields are equal, the queue is empty and processing proceeds to step 2624, where the routine is exited.

If a message is available, processing proceeds to step 2616 where an available buffer is obtained within local memory. The buffer pool for the Shared Memory Driver can be maintained by the operating system in conjunction with the communications handler, as explained below. If a buffer is not available, a wait loop 2617 can be performed. In step 2618, a buffer is obtained and the Dequeued\_offset is used as an offset into the queue to retrieve a pointer to shared memory. The pointer is preferably an offset from the start of the sending sub-Pod's buffer pool. The pointer is used to retrieve the message data from one of the sending sub-Pod's message buffers in shared memory.

In step 2620, the message data is copied to the local buffer. On a NT/UNIX sub-Pod receiving a message from a 2200 operating system, a compaction process can be performed which moves the message bytes into contiguous locations that use all bits (e.g., 64 bits) of a word. This is preferred because 2200 operating system message data occupies only the least-significant four bytes of a word, with the rest being zero-filled. On the 2200 operating system side, the message data can be copied from shared memory using the hardware Block Transfer Unpack (BTU) instruction, which converts message data from 8-bit to 9-bit bytes, and performs little endian (Intel) to big endian (2200 style processor)

- 105 -

conversion. This conversion can be performed in software, firmware, hardware, or any combination thereof.

Alternatively, messages can be stored in shared memory in 2200 style processor  
5 format, whereby a message receiving Intel platform would convert between big and little  
endian and add/remove the extra bit needed by the 2200 style processor.

After the message data is copied to a local buffer, processing proceeds to step 2622,  
where the Shared Memory Driver adds the message to a local memory queue. The Shared  
10 Memory Driver can then check to see that a receiving process (e.g., an application 2208) is  
available to process the message. On the 2200 operating system side, the Shared Memory  
Driver will check to see if a flag indicates that a co-operative processing communications  
program (CPCOMM), developed by Unisys Corporation, is "sleeping." The CPCOMM  
handles communications protocol layers when messages are sent. If CPCOMM is  
15 sleeping, the Share Memory Driver makes a call to the operating system to wake  
CPCOMM up with the newly queued message. Alternatively, polling could be utilized to  
determine if a message is available in local memory.

Figure 27 illustrates an exemplary process for CPCOMM on the 2200 operating  
20 system side that handles receiving messages. As is the case with sending messages,  
CPCOMM does not know that a received message was transferred through shared memory.  
From CPCOMM's point of view, all messages are sent/received over a network.

CPCOMM may be "sleeping" when an interrupt is received from the 2200  
25 operating system. This interrupt is the result of the operating system receiving a call from  
the Shared Memory Driver indicating that a message was queued to CPCOMM's local  
message queue. When CPCOMM is interrupted, it enters a processing loop 2708. The  
process begins at step 2710 where a buffer is acquired in local memory. In step 2712,  
CPCOMM calls the 2200 operating system, passing the buffer address. The 2200  
30 operating system places the buffer in one of the buffer pools associated with one of the  
device drivers, depending on need. The Shared Memory Device Driver is associated with

- 106 -

one of these buffer pools. The buffers in these pools are then available for received message data.

After the buffer address is passed to the operating system, processing proceeds to  
5 step 2714, where CPCOMM checks to see if a message is available on its input queue. Assuming the CPCOMM was interrupted from the operating system, a message is available.

In step 2716, when a message is available, CPCOMM dequeues the message from  
10 its queue, and passes it to the upper layers of the code. Processing then returns to step 2710, where CPCOMM acquires another buffer.

In step 2714, if CPCOMM finds that no more messages are available, processing  
proceeds to step 2718, where CPCOMM determines whether enough empty buffers are  
15 available for use by the various device drivers. If enough buffers are available, processing proceeds to step 2720 where CPCOMM goes to sleep again.

## 20 **V. Conclusions**

It should be understood that embodiments of the present invention can be implemented in hardware, software or a combination thereof. In such embodiments, various components and steps can be implemented in hardware, firmware and/or software  
25 to perform the functions of the present invention. Any presently available or future developed computer software language and/or hardware components can be employed in such embodiments of the present invention. In particular, the pseudo-code discussed and provided above and in the appendixes below can be especially useful for creating the software embodiments.

30

- 107 -

While the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention.

5

- 108 -

## Appendix A

### Forward Windowing Function:

```

5  /* Assign TEMP_ADDR[29:0] to the processors' address PROCESSOR_ADDR[35:6]. Then*/
   /* adjusted for any Relocation, Reclamation, or Shared Windowing, if required. */
   /* if PROCESSOR_ADDR[35:6] ∈ RANGESHARED MEMORY */

   if [SBASEOS] ≤ PROCESSOR_ADDR[35:6] < [TopOfMemoryOS]
10      then TEMP_ADDR[29:0] ← TEMP_ADDR[29:0] + [SBASEMSU - SBASEOS];

   /* if PROCESSOR_ADDR[35:6] ∈ RANGEHIGH MEMORY */

15      elseif [4GB] ≤ [PROCESSOR_ADDR[35:6] < [SBASEOS]

      then TEMP_ADDR[29:0] ← [29:0] + [RLOS - RCOS];

   /* if PROCESSOR_ADDR[35:6] ∈ RANGELOW MEMORY */
20      else

      then TEMP_ADDR[29:0] ← [29:0] + [RLOS];

25  end if;

   /*

```

### Forward Address Translation:

```

30  /* MSU Pair Selection: *****
   /* Inputs: */
   /* TEMP_ADDR: */
   /* Registers used: */
35  /* PAIR_MODE, SMALLEST_PAIR_SZ, PAIR_SEL */
   /* Outputs: */

```

- 109 -

```

/*      TEMP_ADDR:  address after any required adjustment          */
/*      RCVING_PAIR: indicator to the MSU_Pair this address is directed to */
/*****/

5  /* Initialization: *****/
/* TEMP_ADDR[29:0] is the address after any address relocation has been performed */
/*      TEMP_ADDR[29:0] = ADDR_IN[35:6]; */
/* */
/* TOP_OF_INTRLV_RANGE is the address value where there is no more memory left */
10 /* for interleaving, and this is the address where stacking of memory addresses begins. */
/*      TOP_OF_INTRLV_RANGE = 2*SMALLEST_PAIR_SZ; */
/* */
/*****/

15 if PAIR_MODE = INTERLEAVE then
    /* interleave between pairs is enabled */
    /* first check that the address is within the interleaved memory range */

    if (TEMP_ADDR < TOP_OF_INTRLV_RANGE) then
20     /* interleave between Pairs */
        /* use the low order cacheline address bit to select the MSU_Pair# */
        RCVING_PAIR = TEMP_ADDR[0];
        /* then readjust the cacheline address by shifting the cacheline address bits */
        /* to the right by 1 location and zero filling the most significant address bit */
25     TEMP_ADDR = ('0' || (TEMP_ADDR >> 1));

    else
        /* address is over top of interleave range so stack on the larger MSU_Pair */
        RCVING_PAIR = PAIR_SEL;
30     /* adjust address for stacking */
        TEMP_ADDR = (TEMP_ADDR - SMALLEST_PAIR_SZ);

    end if;

35 else [PAIR_MODE = STACK]
    /* stack pairs */
    if (TEMP_ADDR < SMALLEST_PAIR_SZ) then

```

- 110 -

```

/* load into MSU_Pair0 */
    RCVING PAIR = PAIR_SEL;
/* pass address through unadjusted */
    TEMP_ADDR = TEMP_ADDR;
5      else
/* address is outside the address range of MSU_Pair0 */
/* so load overflow (stack) it onto MSU_Pair1 */
    RCVING PAIR not(PAIR_SEL);
/* readjust the address for stacking into the High MSU_Pair */
10      TEMP_ADDR = (TEMP_ADDR - SMALLEST_PAIR_SZ);

    end if;

end if;
15
/* MSU Selection After the Pair is Selected: *****/
/* Inputs: */
/*      TEMP_ADDR - (possibly) adjusted by the previous routines */
/* Registers used: */
20 /*      PAIR0_MODE, PAIR0_SMALLEST_MSU_SZ, PAIR0_SEL */
/* Outputs: */
/*      TEMP_ADDR:  cacheline address after any required adjustment */
/*      RCVING_MSU: indicator to the MSU_# this address is directed to */
/* *****/

25
/* Initialization: *****/
/*
/* PAIR0_TOP_OF_INTLV_RANGE is the address value where no more memory is left */
/* for interleaving between MSUs of MSU_Pair0. */
30 /*      PAIR0_TOP_OF_INTLV_RANGE = (2*PAIR0_SMALLEST_MSU_SZ); */
/*
/* PAIR1_TOP_OF_INTRLV_RANGE is the address value where no more memory is left */
/* for interleaving between MSUs of MSU_Pair1. */
/*      PAIR1_TOP_OF_INTRLV_RANGE = (2*PAIR1_SMALLEST_PAIR_SZ); */
35 /*
/*
*****/

```



- 111 -

```

if RCVING_PAIR = MUS#0 then
    /* Select between MSU#0 or #1 of MSU_Pair0 */
    if PAIRO_MODE = INTERLEAVE then
        /* interleaving between MSU#0 or #1 of MSU_Pair0 is enabled */
5      /* first check that the address is within the interleaved memory range */
        if (TEMP_ADDR < PAIRO_TOP_OF_INTRLV_RANGE) then
            /* use the low order cacheline address bit to select the MSU_Pair# */
            RCVING_MSU = TEMP_ADDR[0];
            /* then readjust the cacheline address by shifting the cacheline address bits */
10          /* to the right by 1 location and zero filling the most significant address bit */
            TEMP_ADDR = ('0' || (TEMP_ADDR >> 1));

        else
            /* stack remainder */
15          /* address is over top of interleave range so stack on the larger MSU */
            RCVING_MSU = PAIRO_SEL;
            /* adjust address for stacking */
            TEMP_ADDR = (TEMP_ADDR - PAIRO_SMALLEST_MSU_SZ);

20        end if;

    else

        /* Stack MSU#0 first then stack the remainder in MSU#1 */
25      if (TEMP_ADDR < PAIRO_SMALLEST_MSU_SZ) then
            /* load into Low MSU, the one designated to receive address '0' */
            RCVING_MSU = PAIRO_SEL;
            /* pass address through unadjusted */
            TEMP_ADDR = TEMP_ADDR;
30          else

            /* load overflow into High MSU */
            RCVING_MSU = not (PAIRO_SEL);
35          /* adjust address for stacking into the High MSU */
            TEMP_ADDR = (TEMP_ADDR - PAIRO_SMALLEST_MSU_SZ);

```

- 112 -

*end if;**end if;*5     *else**if PAIR1\_MODE = INTERLEAVE then**/\* interleaving between MSU#2 or #3 of MSU\_Pair1 is enabled \*/**/\* first check that the address is within the interleaved memory range \*/*

10

*if (TEMP\_ADDR < PAIR1\_TOP\_OF\_INTRLV\_RANGE) then**/\* use the low order cacheline address bit to select the MSU\_Pair# \*/*

RCVING\_MSU = TEMP\_ADDR[0];

*/\* then readjust the cacheline address by shifting the cacheline address bits \*/**/\* to the right by 1 location and zero filling the most significant address bit \*/*

15

TEMP\_ADDR = ('0' || (TEMP\_ADDR &gt;&gt; 1));

*else**/\* stack remainder \*/**/\* address is over top of interleave range so stack on the larger MSU \*/*

20

RCVING\_MSU = PAIR1\_SEL;

*/\* adjust address for stacking \*/*

TEMP\_ADDR = (TEMP\_ADDR - PAIR1\_SMALLEST\_MSU\_SZ);

*end if;*

25

*else**/\* Stack MSU#2 first then stack the remainder in MSU#3 \*/**/\* Note: PAIR1\_SEL = 0 is MSU#2; if = 1 MSU#3 \*/**if (TEMP\_ADDR < PAIR1\_SMALLEST\_MSU\_SZ) then*

30

*/\* load into Low MSU, the one designated to receive address '0' \*/*

RCVING\_MSU = PAIR1\_SEL;

*/\* pass address through unadjusted \*/*

TEMP\_ADDR = TEMP\_ADDR;

*else*

35

*/\* load overflow into High MSU \*/*

RCVING\_MSU = not (PAIR1\_SEL);

- 113 -

/\* adjust address for stacking into the High MSU \*/

TEMP\_ADDR = (TEMP\_ADDR - PAIR1\_SMALLEST\_MSU\_SZ);

end if;

5

end if;

end if;

10

/\*\*\*\*\*

/\* assign MSU\_ADDR[29:0] to the adjusted TEMP\_ADDR[29:0] and \*/

/\* concatenated the RCVING\_PAIR with the RCVING\_MSU indicators to form MSU\_SEL[1:0] \*/

\*\*\*\*\*/

MSU\_ADDR[29:0] = TEMP\_ADDR(29:0);

15

MSU\_SEL[1:0] = (RCVING\_PAIR || RCVING\_MSU);

\*\*\*\*\*/

## Appendix B

### Reverse Translation Algorithm

20

/\* MSU Translation: \*\*\*\*\*/

/\* Inputs: \*/

/\* MSU\_ADDR - \*/

/\* Registers used: \*/

25

/\* PAIR0\_MODE, PAIR0\_SMALLEST\_MSU\_SZ, PAIR0\_SEL \*/

/\* PAIR1\_MODE, PAIR1\_SMALLEST\_MSU\_SZ, PAIR1\_SEL \*/

/\* Outputs: \*/

/\* TEMP\_ADDR: cacheline address after any required adjustment \*/

\*\*\*\*\*/

30

/\* Handling between MSUs of a Pair \*/

if MSU\_PAIR = 1 then

35

/\* MSU\_Pair1 is sending the address back \*/

if PAIR1\_MODE = STACKED then

- 114 -

```

/* MSU#2 and MSU#3 are stacked */

    if (MSU# = PAIR1_SEL then
5         /* addr is in the low MSU */
           TEMP_ADDR = MSU_ADDR;

    else

10         /* addr is in the high MSU, so addr is greater than stacked addr threshold */
           /* stored in PAIR1_SMALLEST_MSU_SZ reg, so adjust addr size */
           TEMP_ADDR = MSU_ADDR + PAIR1_SMALLEST_MSU_SZ;

    end if;

15 else

    /* MSU#2 and MSU#3 are interleaved */
    if (MSU_ADDR > PAIR1_SMALLEST_SMU_SZ) then
20         /* addr is in the overflow MSU, so adjust addr size */
           TEMP_ADDR = MSU_ADDR + PAIR1_SMALLEST_MSU_SZ;

    else

25         /* readjust the cacheline address by shifting the MSU address bits */
           /* to the left by 1 location and replacing the least significant address bit */
           /* with the sending MSU_# (0=MSU#2, 1=MSU#3) */
           TEMP_ADDR = ((MSU_ADDR << 1 || MSU#));

30         end if;

    end if;

    else
35         /* MSU_Pair0 is sending the address back */

        if PAIR0_MODE = STACKED then

```

- 115 -

```

/* MSU#0 and MSU#1 are stacked */

if(MSU# = PAIR0_SEL then
  /* addr is in the low MSU */
5      TEMP_ADDR = MSU_ADDR;

  else

    /* addr is in the high MSU, add back size offset */
10      TEMP_ADDR = MSU_ADDR + PAIR0_SMALLEST_MSU_SZ;

  end if;

  else

15      /* MSU#0 and MSU#1 are interleaved */
      if(MSU_ADDR > PAIR0_SMALLEST_SMU_SZ) then
        /* addr is in the overflow MSU, so adjust addr size */
        TEMP_ADDR = MSU_ADDR + PAIR1_SMALLEST_MSU_SZ;
20      else

        /* readjust the address by shifting the bits to the left by 1 location and */
        /* fill (insert) the least significant address bit with the sending MSU_& */
        /* (0=MSU#0, 1=MSU#1) */
25      TEMP_ADDR = ((MSU_ADDR << 1 || MSU#));

      end if;
    endif;
  endif;

30  /* MSU_Pair Translation: *****/
  /* Inputs: */
  /* TEMP_ADDR - */
  /* Registers used: */
35  /* PAIR_MODE, SMALLEST_MSU_SZ, PAIR_SEL */
  /* Outputs: */
  /* TEMP_ADDR: intermediate address after any required adjustment */

```

- 116 -

```

/* The output is reassigned to the processors' memory request address. It is passed to the */
/* memory relocation procedure for possible mapping adjustment. */
/*     PROCESSOR_ADDR[36:6] = TEMP_ADDR[29:0] */
/*******/
5
/* Handling between pairs */

if PAIR_MODE = 1 then /* MSU_Pair stacking is enabled */

10     if (MSU_Pair = PAIR_SEL) then
        /* addr is in the low Pair */
        TEMP_ADDRS = TEMP_ADDR;

    else
15
        /* addr is in the high Pair so just added the size of the low Pair */
        TEMP_ADDR = TEMP_ADDR + SMALLEST_PAIR_SZ;

    end if;

20
else
    /* MSU_Pair interleaving is enabled */
    if (TEMP_ADDR > SMALLEST_PAIR_SZ) then
        /* addr is in the overflowed Pair */
        TEMP_ADDR = TEMP_ADDR + SMALLEST_PAIR_SZ;
25
    else

        /* readjust the cacheline address by shifting the cacheline address bits */
        /* to the left by 1 location and replacing the least significant address bit */
        /* with the sending MSU_Pair# */
30
        TEMP_ADDR = ((TEMP_ADDR << 1) || MSU_Pair#);

    endif;

35 endif;

```

- 117 -

/\*This ends the Reverse Address Translation Algorithm. Now the address must\*/ /\*be checked to see if it was initially adjusted by the Relocation Function.\*/

### Reverse Windowing Function:

```

5
  /*****/
  /* Adjusted TEMP_ADDR[29:0] to account for mapping around the PCI/APIC      */
  /* range 'hole'. Add back the hole size if relocation was performed. Then assign */
  /* TEMP_ADDR[29:0] to the processors' address PROCESSOR_ADDR[35:6]          */
10 /*****/

  /* if PROCESSOR_ADDR[35:6] ∈ RANGELOW MEMORY */
    if  $[R_L^{OS}] \leq \text{TEMP\_ADDR}[29:0] < [R_L^{OS} + 4\text{GB} - \text{PCI/APIC}_{\text{Hole\_Size}}]$ 
      TEMP_ADDR ← TEMP_ADDR[29:0] -  $[R_L^{OS}]$ ;
15   end if;

  /* if PROCESSOR_ADDR[35:6] ∈ RANGEHIGH MEMORY */
    elsif  $[R_L^{OS} + 4\text{GB} - \text{PCI/APIC}_{\text{Hole\_Size}}] \leq \text{TEMP\_ADDR}[29:0]$ 
      <  $[R_L^{OS} + 4\text{GB} - \text{PCI/APIC}_{\text{Hole\_Size}}]$ 
      then TEMP_ADDR[29:0] ← TEMP_ADDR[29:0] +  $[R_C^{OS} - R_L^{OS}]$ ;
20  /* if PROCESSOR_ADDR[35:6] ∈ RANGESHARED MEMORY */
    elsif  $[S_{\text{BASE}}^{\text{MSU}}] \leq \text{TEMP\_ADDR}[29:0] < [\text{TopOfMemory}^{\text{OS}} + S_{\text{BASE}}^{\text{MSU}} \cdot S_{\text{BASE}}^{\text{OS}}]$ ;
      then TEMP_ADDR[29:0] ← TEMP_ADDR[29:0] +  $[S_{\text{BASE}}^{\text{OS}} - S_{\text{BASE}}^{\text{MSU}}]$ ;
    PROCESSOR_ADDR[35:6] = TEMP_ADDR[29:0];

25 /*****/

```

- 118 -

***What Is Claimed Is:***

1. A computer system comprising:

5 a plurality of processing modules, groups of one or more processing modules being configured as separate partitions within the computer system, each partition operating under the control of a separate operating system;

a main memory to which each processing module is connected such that processor-to-memory latency is the same for each processing module across all of the  
10 main memory, the main memory having defined therein at least one shared memory window to which at least two different partitions have shared access; and

program code, executing on said at least two different partitions, that enables those different partitions to communicate with each other through the shared memory window.

15 2. The computer system recited in claim 1, wherein, for each different partition configured within the computer system, the main memory further has defined therein an exclusive memory window to which only that partition has access.

20 3. The computer system recited in claim 1, wherein ones of the partitions operate under the control of different operating systems.

4. The computer system recited in claim 1, wherein ones of the partitions operate under the control of different instances of a same operating system.

25 5. The computer system recited in claim 1, wherein said program code on a given partition is implemented as one of (i) part of the operating system on that partition, (ii) a separate computer program running on that partition, (iii) a combination of part of the operating system and a separate computer program, (iv) circuitry in which



- 119 -

the program code is hard-code, and (v) program code in combination with circuitry.

6. The computer system recited in claim 1, wherein said program code implements a process by which a sending partition generates an inter-processor  
5 interrupt on a receiving partition to signal the receiving partition that information is being transferred to it through the shared memory window.

7. The computer system recited in claim 6, wherein the shared memory window comprises a set of input queues associated with each partition, each input  
10 queue of the set associated with a given partition corresponding to another partition and storing entries representing communications from that other partition.

8. The computer system recited in claim 7, wherein the shared memory window further comprises a plurality of pages of memory that can be allocated to the  
15 partitions, as needed, to facilitate communication of information between them.

9. The computer system recited in claim 8, wherein an input queue entry representing a communication between a sending partition and a receiving partition may comprise a handle to one or more pages of the shared memory window.  
20

10. The computer system recited in claim 8, wherein a sending partition may use one or more allocated pages to store data representing a message to be communicated to a receiving partition.

25 11. The computer system recited in claim 7, wherein in order for one partition (a sending partition) to communicate with another partition (a receiving partition), the program code on the sending partition:

(i) causes an entry to be created in the input queue of the receiving partition that corresponds to the sending partition; and

30 (ii) causes an inter-processor interrupt to be generated on the receiving partition to signal the receiving partition that the entry has been created in that input queue.

- 120 -

12. The computer system recited in claim 11, wherein when the inter-processor interrupt is detected on the receiving partition, the program code on the receiving partition:

- 5 (i) causes each of its input queues to be examined to determine which of the input queues contain entries representing communications from other partitions; and
- (ii) causes any such entries to be extracted from the input queues that contain them.

10 13. The computer system recited in claim 2, wherein the shared memory window comprises a set of input queues associated with each partition, each input queue of the set associated with a given partition corresponding to another partition and storing entries representing communications from that other partition, each input queue capable of storing a pre-defined number of entries and containing an overflow flag that

15 is caused to be set whenever the input queue is full.

14. The computer system recited in claim 13, wherein in order for one partition (a sending partition) to communicate with another partition (a receiving partition), the program code on the sending partition:

- 20 (i) causes an entry to be created in the input queue of the receiving partition that corresponds to the sending partition;
- (ii) causes the overflow flag of the input queue to be set if the creation of the entry causes the input queue to become full; and
- (iii) causes an inter-processor interrupt to be generated on the receiving
- 25 partition to signal the receiving partition that the entry has been created in that input queue.

15. The computer system recited in claim 14, wherein when an inter-processor interrupt is detected on the receiving partition, the program code on the

30 receiving partition:

- (i) causes each of its input queues to be examined to determine which of the

- 121 -

input queues contain entries representing communications from other partitions;

(ii) causes any such entries to be extracted from the input queues that contain them; and

(iii) if the overflow flag of any input queue from which entries were extracted  
5 was set, causes the overflow flag to be reset.

16. The computer system recited in claim 15, wherein each partition maintains a location in its exclusive memory window for storing input queue entries that could not be placed in a designated input queue because the overflow flag of that  
10 input queue was set previously to indicate that the input queue became full, the entries stored in the exclusive memory window location remaining there until the overflow flag of the designated input queue is reset indicating that it is no longer full, whereupon the entries are transferred to the input queue.

15 17. The computer system recited in claim 15, wherein each partition maintains a location in its exclusive memory window in which queue entries extracted from a given input queue are stored temporarily.

18. The computer system recited in claim 15, wherein each partition assigns  
20 a separate interrupt vector for each other partition from which it may receive an inter-processor interrupt, and wherein the sending partition specifies the interrupt vector assigned to it when sending an inter-processor interrupt to the receiving partition, and further wherein, in response to the inter-processor interrupt, the receiving partition:

(i') uses the specified interrupt vector to identify the input queue associated with  
25 the sending partition;

(ii') causes any entries to be extracted from the identified input queue; and

(iii') if the overflow flag of the identified input queue was set, causes the overflow flag to be reset.

30 19. The computer system recited in claim 7, wherein each input queue contains a count indicating the number of entries stored therein.

- 122 -

20. The computer system recited in claim 8, wherein the shared memory window further comprises a table indicating, for each page, whether the page is in-use or is available for allocation.

5

21. The computer system recited in claim 20, wherein the pages that are available for allocation are linked together to form a linked-list of available pages.

22. The computer system recited in claim 8, wherein each partition may have ownership rights in a particular page, and wherein the page has a header containing information that specifies which partitions have ownership rights in the page.

23. The computer system recited in claim 22, wherein the header of the page further comprises a lock field by which one partition may acquire exclusive access to a page in order to update ownership information in the header of the page, thereby providing a mechanism to synchronize multiple accesses to the page by different partitions.

24. The computer system recited in claim 23, wherein the shared memory window has a system-wide lock field associated with it by which one partition may acquire exclusive access to the shared memory pages in order to allocate one or more pages of the shared memory window, thereby providing a mechanism to synchronize multiple requests for allocation of memory pages by different partitions.

25

25. The computer system recited in claim 24, wherein ownership information of a page may be updated by acquiring the lock field of the page, without having to acquire the system-wide lock field.

30

26. The computer system recited in claim 1, wherein the program code

- 123 -

implements a polling process by which each partition polls an area within the shared memory window to determine whether any communications intended for it have been placed in the shared memory window by another partition.

5           27.     The computer system recited in claim 26, wherein each partition is allocated a separate pool of message buffers in which it may place communications intended for other partitions.

10           28.     The computer system recited in claim 26, wherein the area comprises a plurality of output queues, one for each partition, the output queue for a given partition indicating whether that partition has placed in the shared memory window any communications intended for any of the other partitions, each partition polling the output queues of the other partitions to determine whether those other partitions have placed any communications intended for it in the shared memory window.

15           29.     The computer system recited in claim 28, wherein for any communications placed in the shared memory window by a sending partition and intended to be received by another partition, the output queue of the sending partition specifies the location within the shared memory window of a buffer containing that communication.

20           30.     The computer system recited in claim 28, wherein the output queue of a given partition comprises one or more node-to-node queues, one associated with each other partition to which it may pass communications, each node-to-node queue indicating whether communications intended for the partition with which it is associated have been placed in the shared memory window, each partition polling the node-to-node queues associated with it in the output queues of each other partition to determine whether any of those other partitions have placed any communications intended for it in the shared memory window.

30

31.     The computer system recited in claim 30, wherein for any

- 124 -

communications placed in the shared memory window by a sending partition and intended to be received by a receiving partition, the node-to-node queue associated with the receiving partition in the output queue of the sending partition indicates the existence of that communication and specifies the location within the shared memory window of a buffer containing that communication.

32. The computer system recited in claim 29, wherein each partition of the computer system further comprises a shared memory driver that receives a request to send a communication to another partition and, in response thereto, causes the communication to be placed in an available buffer in the shared memory window and causes an indication of the communication to be placed in the output queue of the sending partition.

33. The computer system recited in claim 32, wherein the shared memory driver on each partition implements a same interface as a network device driver to enable application programs and the operating system on that partition to send communications to other partitions via the shared memory window in the same manner that communications are sent to other computer systems over a network, the shared memory driver thereby providing the appearance that communications between partitions via shared memory are being made over a network.

34. A method for use in a computer system comprising a plurality of processing modules and a main memory to which each processing module is connected such that processor-to-memory latency is the same for each processing module across all of the main memory, said method comprising:

configuring different processing modules of the computer system into different partitions within the computer system, each partition comprising one or more processing modules and each partition executing a separate operating system;  
assigning at least one shared memory window to which a plurality of the partitions have shared access;  
communicating information from a sending partition to a receiving partition by the

- 125 -

sending partition storing the information in a location within the shared memory window and by the sending partition causing an inter-processor interrupt to be generated on the receiving partition to signal the receiving partition that the information to be communicated to it is stored in that location.

5

35. The method recited in claim 34, further comprising the step of creating in the shared memory window, for each partition, a set of input queues, each input queue of the set of a given partition corresponding to another partition and storing entries representing communications from that other partition.

10

36. The method recited in claim 35, wherein said communicating step comprises the steps of:

creating an entry in the input queue of the receiving partition that corresponds to the sending partition, the entry representing a communication to be sent to the receiving partition; and

15

generating an inter-processor interrupt on the receiving partition to signal the receiving partition that the entry has been created.

37. The method recited in claim 36, further comprising the following steps performed on the receiving partition:

20

detecting the inter-processor interrupt;

examining, in response to the inter-processor interrupt, each of the set of input queues of the receiving partition to determine which of the input queues contains entries representing communications from other partitions; and

25

extracting any such entries from the input queues that contain them.

38. The method recited in claim 37, wherein each partition assigns a separate interrupt vector for each other partition from which it may receive an inter-processor interrupt, and wherein the sending partition specifies the interrupt vector assigned to it by the receiving partition when sending an inter-processor interrupt to the receiving partition, and further wherein, in response to the inter-processor interrupt, the

30

- 126 -

receiving partition performs the following alternative steps:

identifying the input queue associated with the sending partition using the specified interrupt vector;

extracting any entries from the identified input queue; and

5 if the overflow flag of the identified input queue was set, resetting the overflow flag.

39. The method recited in claim 37, further comprising the steps of:

10 defining, for each partition, an exclusive memory window within the main memory to which the partition has exclusive access; and

after extracting an entry from an input queue of a receiving partition, storing the entry in a temporary location within its exclusive memory window.

40. The method recited in claim 35, further comprising the step of defining a  
15 region within the shared memory window comprising a plurality of pages of memory that can be allocated to the partitions, as needed, to facilitate communication between them.

41. The method recited in claim 40, wherein said communicating step  
20 comprises the steps of:

allocating one or more pages of the shared memory window to the sending partition;

copying information to be communicated from the sending partition to the receiving partition into the allocated pages;

creating an entry in the input queue of the receiving partition that corresponds to the sending partition, the entry containing a handle to the allocated pages; and

generating an inter-processor interrupt on the receiving partition to signal the receiving partition that the entry has been created.

30 42. The method recited in claim 41, further comprising the following steps performed on the receiving partition:



- 127 -

detecting the inter-processor interrupt;

examining, in response to the inter-processor interrupt, each input queue of the set of the receiving partition to determine which of the input queues contains entries representing communications from other partitions;

- 5        extracting any such entries from the input queues that contain them; and  
if any extracted entry contains a handle to one or more allocated pages of memory, utilizing the handle to access the allocated pages.

43.     The method recited in claim 41, wherein each partition assigns a  
10    separate interrupt vector for each other partition from which it may receive an inter-processor interrupt, and wherein the sending partition specifies the interrupt vector assigned to it by the receiving partition when sending an inter-processor interrupt to the receiving partition, and further wherein, in response to the inter-processor interrupt, the receiving partition performs the following alternative steps:

- 15        identifying the input queue associated with the sending partition using the specified interrupt vector;  
extracting any entries from the identified input queue; and  
if any extracted entry contains a handle to one or more allocated pages of memory, utilizing the handle to access the allocated pages.

20

44.     The method recited in claim 35, wherein each input queue is capable of storing a pre-defined number of entries and contains an overflow flag that can be set whenever the input queue is full.

- 25        45.     The method recited in claim 44, wherein said communicating step comprises the steps of:

creating an entry in the input queue of the receiving partition that corresponds to the sending partition, the entry representing information to be communicated to the receiving partition;

- 30        setting the overflow flag of the input queue if the creation of the entry causes the input queue to become full; and

- 128 -

generating an inter-processor interrupt on the receiving partition to signal the receiving partition that the entry has been created in that input queue.

46. The method recited in claim 45, further comprising the following steps  
5 performed on the receiving partition:

detecting the inter-processor interrupt;

examining, in response to the inter-processor interrupt, each of the set of input  
queues of the receiving partition to determine which of the input queues contains  
entries representing communications from other partitions;

10 extracting any such entries from the input queues that contain them; and,  
if the overflow flag of any input queue from which entries were extracted was  
set, resetting the overflow flag.

47. The method recited in claim 45, wherein each partition assigns a  
15 separate interrupt vector for each other partition from which it may receive an inter-  
processor interrupt, and wherein the sending partition specifies the interrupt vector  
assigned to it by the receiving partition when sending an inter-processor interrupt to the  
receiving partition, and further wherein, in response to the inter-processor interrupt, the  
receiving partition performs the following alternative steps:

20 identifying the input queue associated with the sending partition using the  
specified interrupt vector;

extracting any entries from the identified input queue; and

if the overflow flag of any input queue from which entries were extracted was  
set, resetting the overflow flag.

25

48. The method recited in claim 35, further comprising the step of defining  
for each partition, a separate exclusive memory window within the main memory to  
which the partition has exclusive access.

30 49. The method recited in claim 48, wherein each input queue is capable of  
storing a pre-defined number of entries and contains an overflow flag that can be set

- 129 -

whenever the input queue is full.

50. The method recited in claim 49, wherein said communicating step comprises the steps of:

5 creating an entry in the input queue of the receiving partition that corresponds to the sending partition, the entry representing information to be communicated to the receiving partition;

setting the overflow flag of the input queue if the creation of the entry causes the input queue to become full, and thereafter temporarily storing any additional entries  
10 to be created in that input queue in a location within the exclusive memory window of the sending partition so that they can later be placed in the input queue when it is no longer full; and

generating an inter-processor interrupt on the receiving partition to signal the receiving partition that the input queue contains entries for the receiving partition.  
15

51. The method recited in claim 50, further comprising the following steps performed on the receiving partition:

detecting the inter-processor interrupt and locating an input queue containing entries representing communications from another partition;

20 extracting the entries from the input queue; and,

if the overflow flag of any input queue from which entries were extracted was set, resetting the overflow flag and sending a communication back to the corresponding sending partition, via an entry in the corresponding input queue of that sending partition, indicating that that sending partition can resume sending entries that may  
25 have been temporarily stored in its exclusive memory window.

52. A method for allowing a plurality of operating systems to operate on a computer system, wherein the computer system comprises a plurality of processing modules and a main memory, said method comprising:

30 configuring different processing modules of the computer system into different partitions within the computer system, each partition comprising one or more processing

- 130 -

modules and each partition executing a separate operating system;

assigning each partition an exclusive memory window within the main memory and assigning at least one shared memory to which a plurality of the partitions have shared access;

5 communicating messages from one partition to another partition via the shared memory in accordance with a polling process by which each partition polls an area of the shared memory to determine whether messages intended for it have been placed in the shared memory window by another partition.

10 53. The method recited in claim 52, further comprising the step of allocating, to each of the separate operating systems, a separate pool of message buffers within the shared memory window in which the partition may place messages intended for other partitions.

15 54. The method recited in claim 53, wherein the area of the shared memory that is polled by the partitions comprises a plurality of output queues, one for each partition, the output queue for a given partition indicating whether that partition has placed in the shared memory window any messages intended for any of the other partitions, said communicating step comprising each partition polling the output queues  
20 of the other partitions to determine whether those other partitions have placed any messages intended for it in the shared memory window.

55. The method recited in claim 54, further comprising the steps of:  
a partition intending to send a message to another partition (sending partition)  
25 placing the message in a buffer within the shared memory window;

the sending partition storing in its output queue an indication that a message intended for the other partition has been placed in the buffer and in addition storing a pointer to the buffer;

the partition to which the message is intended polling the output queue of the  
30 sending partition, determining therefrom that the message has been placed in the buffer,

- 131 -

and using the stored pointer to locate the buffer and to retrieve the message from the buffer.

56. A computer system comprising:

5 a plurality of processing modules, each processing module comprising a plurality of processors, groups of one or more processing modules being configured as separate partitions within the computer system, each partition operating under the control of a separate operating system;

10 a main memory within which each partition is assigned an exclusive memory window to which only that partition has access and in which the operating system of that partition operates; and

15 means for mapping the physical address space of the processors in each partition to the respective exclusive memory window assigned to the partition, whereby the exclusive memory windows of each partition are made to appear to their respective operating systems as having a same base physical address in the main memory.

57. The computer system recited in 56, wherein each exclusive memory window is made to appear to its respective operating systems as having a base physical address of zero.

20

58. The computer system recited in claim 56, wherein said means for mapping comprises means for relocating a reference to a location within the physical address space of the processors on a given partition to the corresponding location within the exclusive memory window assigned to that partition.

25

59. The computer system recited in claim 58, wherein said means for relocating comprises:

a register that holds an offset ( $R_L^{OS}$ ) from the base physical address of main memory to the start of the exclusive memory window assigned to said given partition;

30 and

an adder for adding the offset ( $R_L^{OS}$ ) to each reference by a processor in said

- 132 -

given partition to a location within its physical address space, thereby relocating those references to their corresponding locations within the exclusive memory window.

60. The computer system recited in claim 56, wherein the physical address  
5 space of the processors of a given partition may contain a range of addresses  
unavailable for memory storage, the unavailable range defining a memory hole,  
addresses above the memory hole defining a high memory range and addresses below  
the memory hole defining a low memory range, the computer system further  
comprising means for reclaiming for other uses that portion of the exclusive memory  
10 window of said given partition that would otherwise correspond to the memory hole as  
a result of said mapping function.

61. The computer system recited in claim 60, wherein said means for  
mapping comprises:  
15 a register that holds an offset ( $R_L^{OS}$ ) from the base physical address of main  
memory to the start of the exclusive memory window assigned to said given partition;  
a register that holds an a value ( $R_C^{OS}$ ) representing the size of the memory hole;  
and an adder for:  
(i) adding the offset ( $R_L^{OS}$ ) to each reference by a processor in said given  
20 partition to a location within the low memory range of its physical address space,  
thereby relocating those references to their corresponding locations within the exclusive  
memory window, and  
(ii) adding the offset minus the value representing the size of the memory hole  
( $R_L^{OS} - R_C^{OS}$ ) to each reference by a processor in said given partition to a location within  
25 the high memory range of its physical address space, thereby relocating those  
references to their corresponding locations within the exclusive memory window and  
reclaiming that portion of the exclusive memory window that would otherwise have  
corresponded to the memory hole.

30 62. The computer system recited in claim 56, wherein the main memory  
further comprises a shared memory window separate from the exclusive memory

- 133 -

windows, and wherein the computer system further comprises means for mapping, on each partition, a designated portion of the physical address space of the processors of that partition to the shared memory window within the main memory.

5           63.     The computer system recited in claim 59, wherein said means for mapping a designated portion of the physical address space of the processors on a given partition to the shared memory window comprises:

              a register that holds an offset ( $S_{\text{BASE}}^{\text{OS}}$ ) from the base address of the physical address space of the processors on the partition to the start of said designated portion of  
10   that physical address space;

              a register that holds an offset ( $S_{\text{BASE}}^{\text{MSU}}$ ) from the base address of the main memory to the start of the shared memory window within the main memory; and

              an adder that adds the difference between the offsets ( $S_{\text{BASE}}^{\text{MSU}} - S_{\text{BASE}}^{\text{OS}}$ ) to each  
15   reference by a processor in said given partition to a location within said designated portion, thereby relocating those references to their corresponding locations within the shared memory window of the main memory.

              64.     The computer system recited in claim 56, wherein ones of the partitions operate under the control of different operating systems.

20

              65.     The computer system recited in claim 56, wherein ones of the partitions operate under the control of different instances of a same operating system.

              66.     In a computer system comprising (i) a plurality of processing modules,  
25   each processing module comprising a plurality of processors, groups of one or more processing modules being configured as separate partitions within the computer system, each partition operating under the control of a separate operating system, and (ii) a main memory within which each partition is assigned an exclusive memory window to which only that partition has access and in which the operating system of that partition  
30   operates, a method for making the exclusive memory windows of each partition appear to their respective operating systems as having a same base physical address in the

- 134 -

main memory, said method comprising relocating a reference to a location within the physical address space of the processors on a given partition to the corresponding location within the exclusive memory window assigned to that partition.

5           67.     The method recited in claim 66, wherein said relocating step comprises:  
              storing a value representing an offset ( $R_L^{OS}$ ) from the base physical address of  
              the main memory to the start of the exclusive memory window assigned to a given  
              partition; and

              adding the offset ( $R_L^{OS}$ ) to each reference by a processor in that partition to a  
10           location within its physical address space, thereby relocating those references to their  
              corresponding locations within the exclusive memory window.

              68.     The method recited in claim 56, wherein the physical address space of  
              the processors of a given partition may contain a range of addresses unavailable for  
15           memory storage, the unavailable range defining a memory hole, addresses above the  
              memory hole defining a high memory range and addresses below the memory hole  
              defining a low memory range, said method further comprising reclaiming for other uses  
              that portion of the exclusive memory window of said given partition that would  
              otherwise correspond to the memory hole as a result of said relocating step.

20

              69.     The method recited in claim 68, wherein said relocating and reclaiming  
              steps comprise:

              storing a value representing an offset ( $R_L^{OS}$ ) from the base physical address of  
              main memory to the start of the exclusive memory window assigned to said given  
25           partition;

              storing a value ( $R_C^{OS}$ ) representing the size of the memory hole;  
              adding the offset ( $R_L^{OS}$ ) to each reference by a processor in said given partition  
              to a location within the low memory range of its physical address space, thereby  
              relocating those references to their corresponding locations within the exclusive  
30           memory window; and

              adding the offset minus the size of the memory hole ( $R_L^{OS} - R_C^{OS}$ ) to each



- 135 -

reference by a processor in said given partition to a location within the high memory range of its physical address space, thereby relocating those references to their corresponding locations within the exclusive memory window and reclaiming that portion of the exclusive memory window that would otherwise have corresponded to the memory hole.

70. The method recited in claim 56, wherein the main memory further comprises a shared memory window separate from the exclusive memory windows, and wherein said method further comprises:

10 designating, on each partition, a portion of the physical address space of the processors of that partition to correspond to the shared memory window within the main memory; and

relocating any reference by a processor of a partition to a location within the designated portion of its physical address space to the corresponding location within the shared memory window within the main memory.

71. The method recited in claim 70, wherein said step of relocating a reference by a processor on a partition to the designated portion of its physical address to the corresponding location in the shared memory window, comprises:

20 storing a value representing an offset ( $S_{\text{BASE}}^{\text{OS}}$ ) from the base address of the physical address space of the processor on that partition to the start of said designated portion of that physical address space;

storing a value representing an offset ( $S_{\text{BASE}}^{\text{MSU}}$ ) from the base address of the main memory to the start of the shared memory window within the main memory; and

25 adding the difference between the stored offsets ( $S_{\text{BASE}}^{\text{MSU}} - S_{\text{BASE}}^{\text{OS}}$ ) to any reference by a processor in that partition to a location within the designated portion, thereby relocating those references to their corresponding locations within the shared memory window of the main memory.

30 72. The method recited in 56, wherein each exclusive memory window is made to appear to its respective operating system as having a base physical address of

- 136 -

zero.

73. A computer system comprising:
- 5 a plurality of processing modules, each processing module comprising a plurality of processors, groups of one or more processing modules being configured as separate partitions within the computer system, each partition operating under the control of a separate operating system;
- 10 a main memory within which each partition is assigned an exclusive memory window to which only that partition has access and in which the operating system of that partition operates, and within which there is also defined a shared memory window to which a plurality of partitions have shared access;
- means for mapping the physical address space of the processors in each partition to the respective exclusive memory window assigned to the partition, whereby the exclusive memory windows of each partition are made to appear to their respective
- 15 operating systems as having a same base physical address in the main memory; and
- program code, executing on said plurality of partitions, that enables those partitions to communicate with each other through the shared memory window.

74. The computer system recited in claim 73, wherein ones of the partitions
- 20 operate under the control of different operating systems.

75. The computer system recited in claim 73, wherein ones of the partitions operate under the control of different instances of a same operating system.

- 25 76. The computer system recited in claim 73, wherein said program code implements a process by which a sending partition generates an inter-processor interrupt on a receiving partition to signal the receiving partition that information is being transferred to it through the shared memory window.

- 30 77. The computer system recited in claim 76, wherein the shared memory window comprises a set of input queues associated with each partition, each input

- 137 -

queue of the set associated with a given partition corresponding to another partition and storing entries representing communications from that other partition.

78. The computer system recited in claim 77, wherein the shared memory  
5 window further comprises a plurality of pages of memory that can be allocated to the partitions, as needed, to facilitate communication of information between them.

79. The computer system recited in claim 78, wherein each partition may  
10 have ownership rights in a particular page, and wherein the page has a header containing information that specifies which partitions have ownership rights in the page.

80. The computer system recited in claim 79, wherein the header of the page  
15 further comprises a lock field by which one partition may acquire exclusive access to a page in order to update ownership information in the header of the page, thereby providing a mechanism to synchronize multiple accesses to the page by different partitions.

81. The computer system recited in claim 80, wherein the shared memory  
20 window has a system-wide lock field associated with it by which one partition may acquire exclusive access to the shared memory pages in order to allocate one or more pages, thereby providing a mechanism to synchronize multiple requests for allocation of memory pages by different partitions.

82. The computer system recited in claim 80, wherein the ownership of a  
25 page can be updated by acquiring the lock field of that page, without having to acquire the system-wide lock field.

83. The computer system recited in claim 77, wherein in order for one  
30 partition (a sending partition) to communicate with another partition (a receiving partition), the program code on the sending partition:

- 138 -

(i) causes an entry to be created in the input queue of the receiving partition that corresponds to the sending partition; and

(ii) causes an inter-processor interrupt to be generated on the receiving partition to signal the receiving partition that the entry has been created in that input queue.

5

84. The computer system recited in claim 83, wherein when the inter-processor interrupt is detected on the receiving partition, the program code on the receiving partition:

(i) causes each of its input queues to be examined to determine which of the  
10 input queues contain entries representing communications from other partitions; and

(ii) causes any such entries to be extracted from the input queues that contain them.

85. The computer system recited in claim 77, wherein each input queue is  
15 capable of storing a pre-defined number of entries and contains an overflow flag that is caused to be set by a sending partition whenever the input queue becomes full, and which is reset by a receiving partition whenever entries are extracted from the input queue.

20 86. The computer system recited in claim 73, wherein the program code implements a polling process by which each partition polls an area within the shared memory window to determine whether any communications intended for it have been placed in the shared memory window by another partition.

25 87. The computer system recited in claim 86, wherein the area comprises a plurality of output queues, one for each partition, the output queue for a given partition indicating whether that partition has placed in the shared memory window any communications intended for any of the other partitions, each partition polling the output queues of the other partitions to determine whether those other partitions have  
30 placed any communications intended for it in the shared memory window.

- 139 -

88. The computer system recited in claim 87, wherein for any communications placed in the shared memory window by a sending partition and intended to be received by another partition, the output queue of the sending partition specifies the location within the shared memory window of a buffer containing that communication.

89. The computer system recited in claim 88, wherein each partition is allocated a separate pool of message buffers in which it may place communications intended for other partitions.

10

90. The computer system recited in 73, wherein each exclusive memory window is made to appear to its respective operating systems as having a base physical address of zero.

15 91. The computer system recited in claim 73, wherein said means for mapping comprises means for relocating a reference to a location within the physical address space of the processors on a given partition to the corresponding location within the exclusive memory window assigned to that partition.

20 92. The computer system recited in claim 91, wherein said means for relocating comprises:

a register that holds an offset ( $R_L^{OS}$ ) from the base physical address of main memory to the start of the exclusive memory window assigned to said given partition; and

25 an adder for adding the offset ( $R_L^{OS}$ ) to each reference by a processor in said given partition to a location within its physical address space, thereby relocating those references to their corresponding locations within the exclusive memory window.

93. The computer system recited in claim 73, wherein the physical address space of the processors of a given partition may contain a range of addresses unavailable for memory storage, the unavailable range defining a memory hole,

30

- 140 -

addresses above the memory hole defining a high memory range and addresses below the memory hole defining a low memory range, the computer system further comprising means for reclaiming for other uses that portion of the exclusive memory window of said given partition that would otherwise correspond to the memory hole as  
5 a result of said mapping function.

94. The computer system recited in claim 93, wherein said means for mapping comprises:

a register that holds an offset ( $R_L^{OS}$ ) from the base physical address of main  
10 memory to the start of the exclusive memory window assigned to said given partition;  
a register that holds an a value ( $R_C^{OS}$ ) representing the size of the memory hole;  
and an adder for:

(i) adding the offset ( $R_L^{OS}$ ) to each reference by a processor in said given partition to a location within the low memory range of its physical address space,  
15 thereby relocating those references to their corresponding locations within the exclusive memory window, and

(ii) adding the offset minus the value representing the size of the memory hole ( $R_L^{OS} - R_C^{OS}$ ) to each reference by a processor in said given partition to a location within the high memory range of its physical address space, thereby relocating those  
20 references to their corresponding locations within the exclusive memory window and reclaiming that portion of the exclusive memory window that would otherwise have corresponded to the memory hole.

95. The computer system recited in claim 73, further comprising means for  
25 mapping, on each partition, a designated portion of the physical address space of the processors of that partition to the shared memory window within the main memory.

96. The computer system recited in claim 95, wherein said means for mapping a designated portion of the physical address space of the processors on a given  
30 partition to the shared memory window comprises:

a register that holds an offset ( $S_{BASE}^{OS}$ ) from the base address of the physical

- 141 -

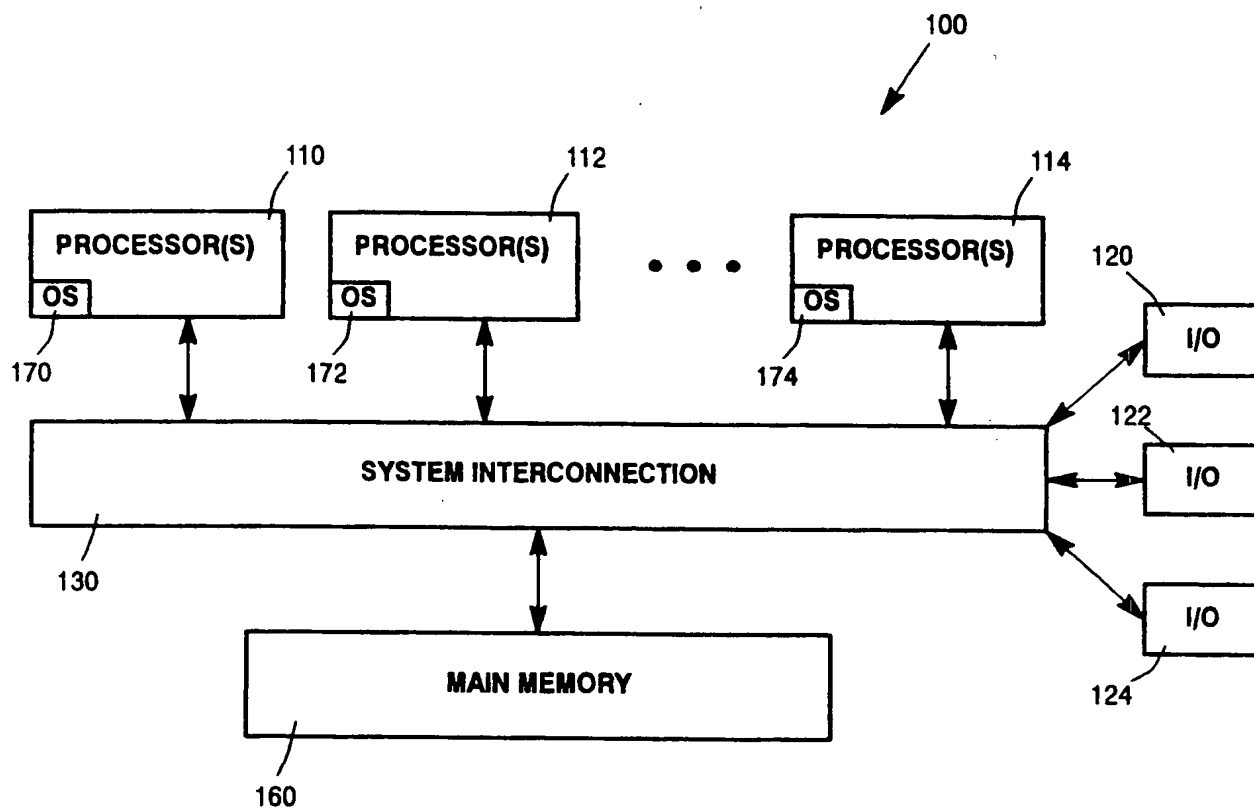
address space of the processors on the partition to the start of said designated portion of that physical address space;

a register that holds an offset ( $S_{\text{BASE}}^{\text{MSU}}$ ) from the base address of the main memory to the start of the shared memory window within the main memory; and

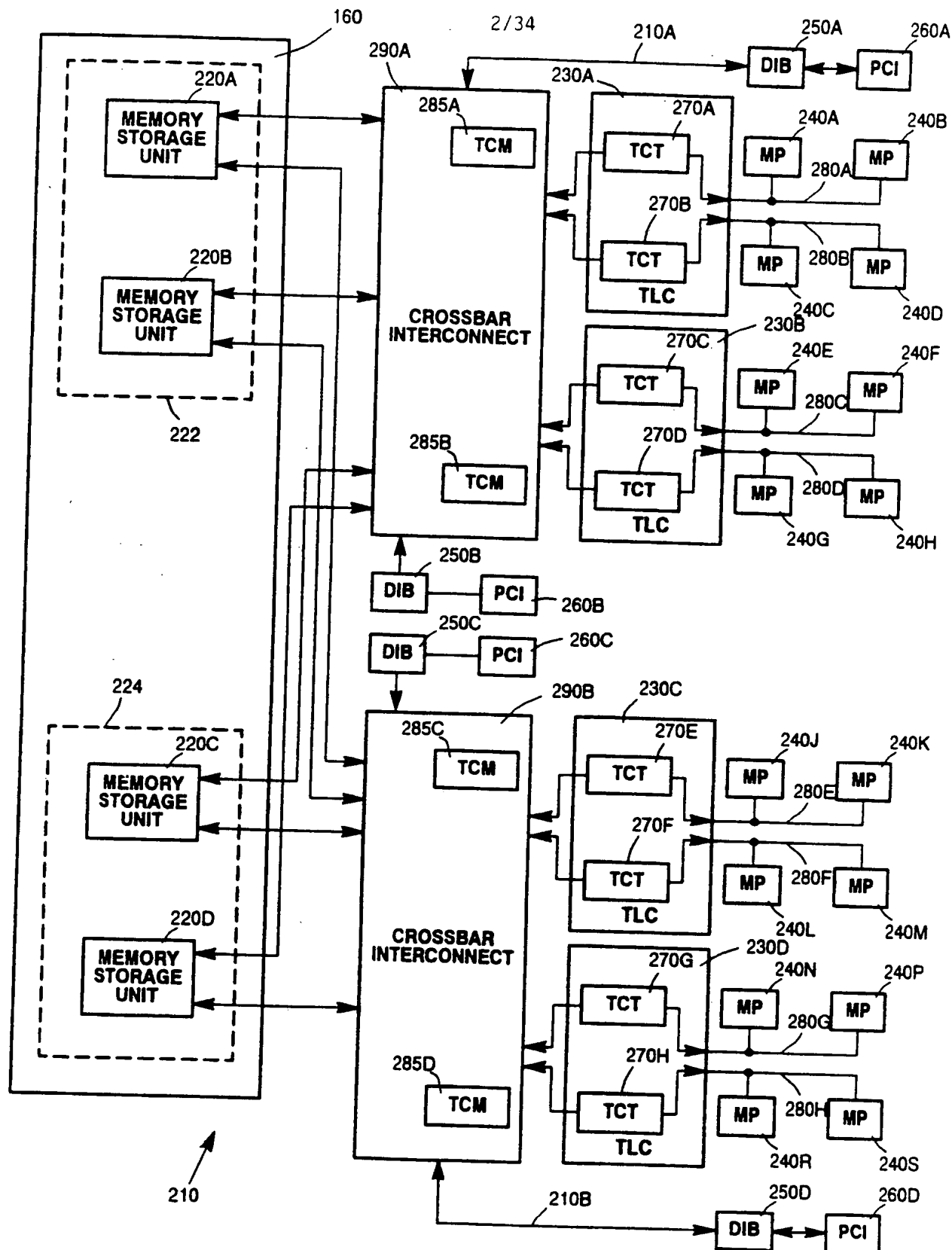
- 5 an adder that adds the difference between the offsets ( $S_{\text{BASE}}^{\text{MSU}} - S_{\text{BASE}}^{\text{OS}}$ ) to each reference by a processor in said given partition to a location within said designated portion, thereby relocating those references to their corresponding locations within the shared memory window of the main memory.

10

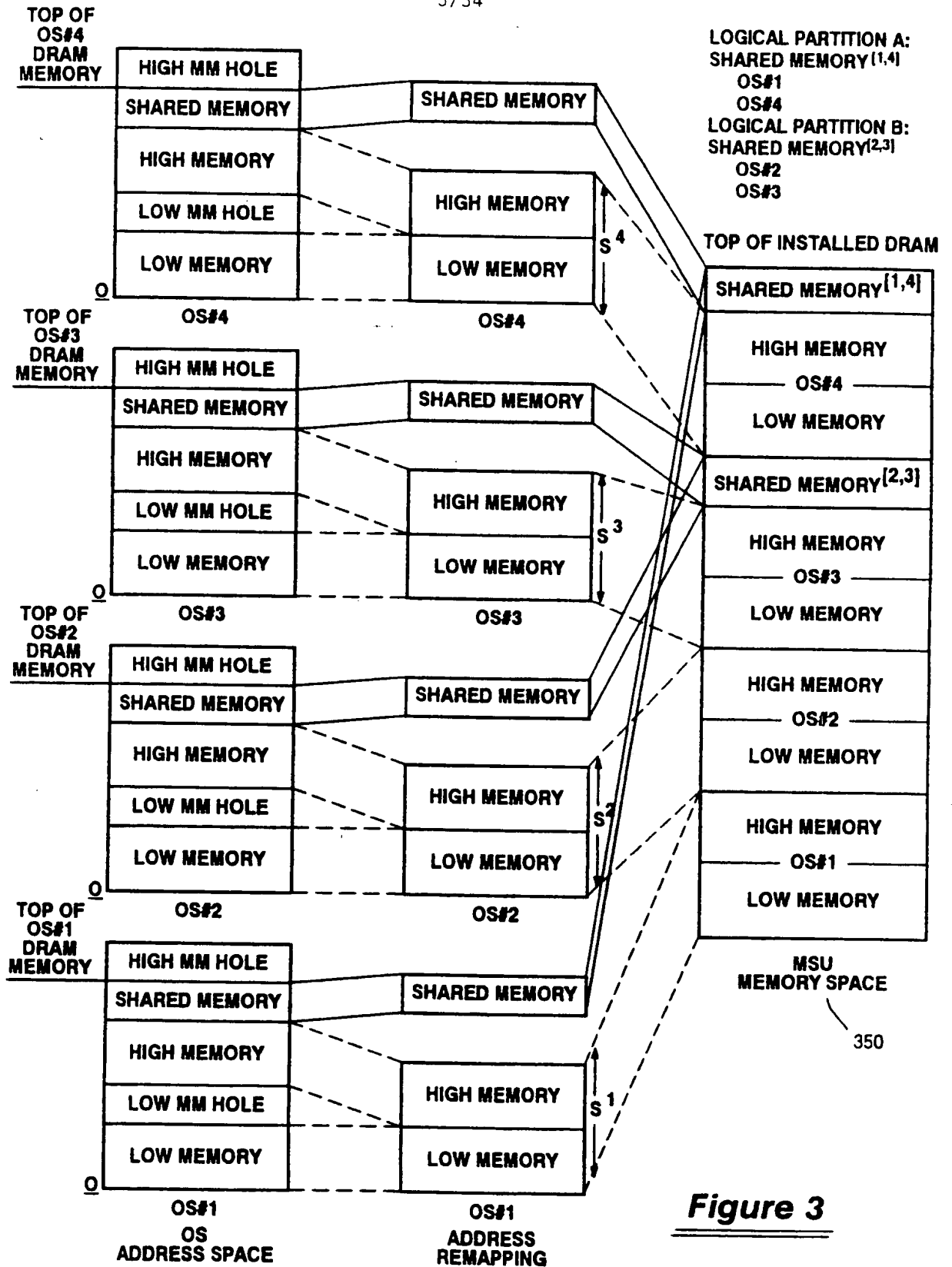
1/34

**Figure 1**

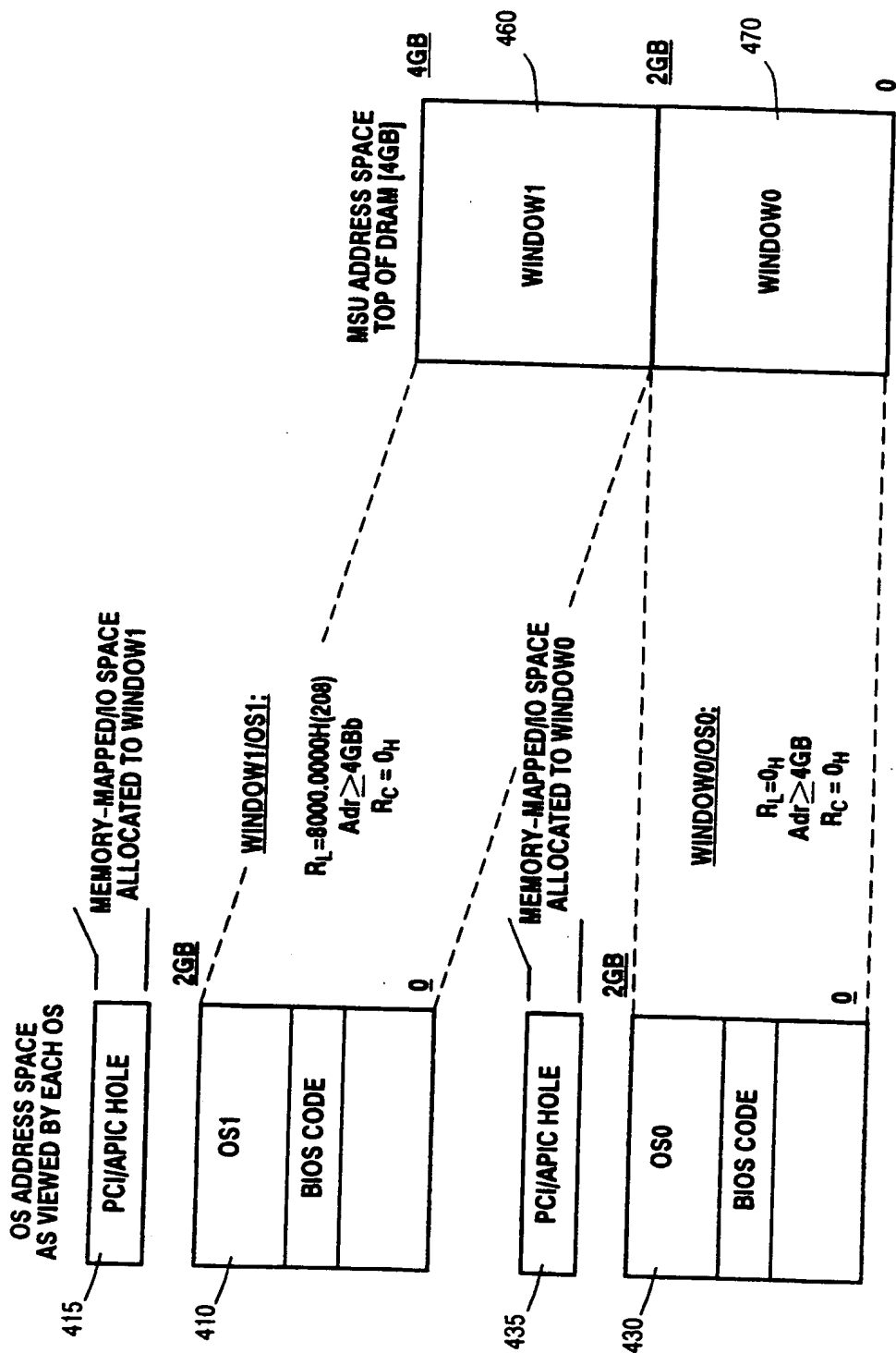


**Figure 2**

3/34

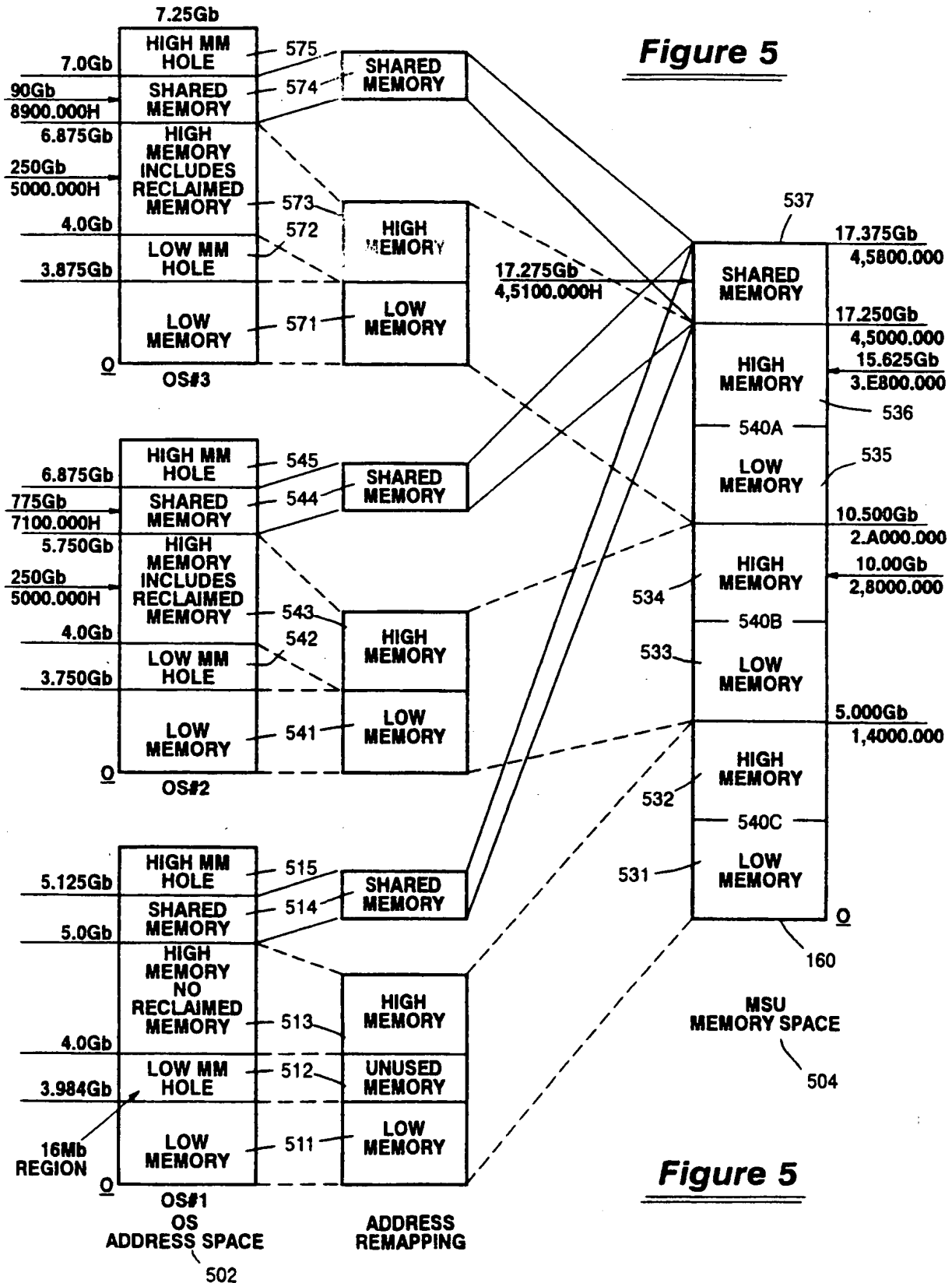
**Figure 3**

4/34

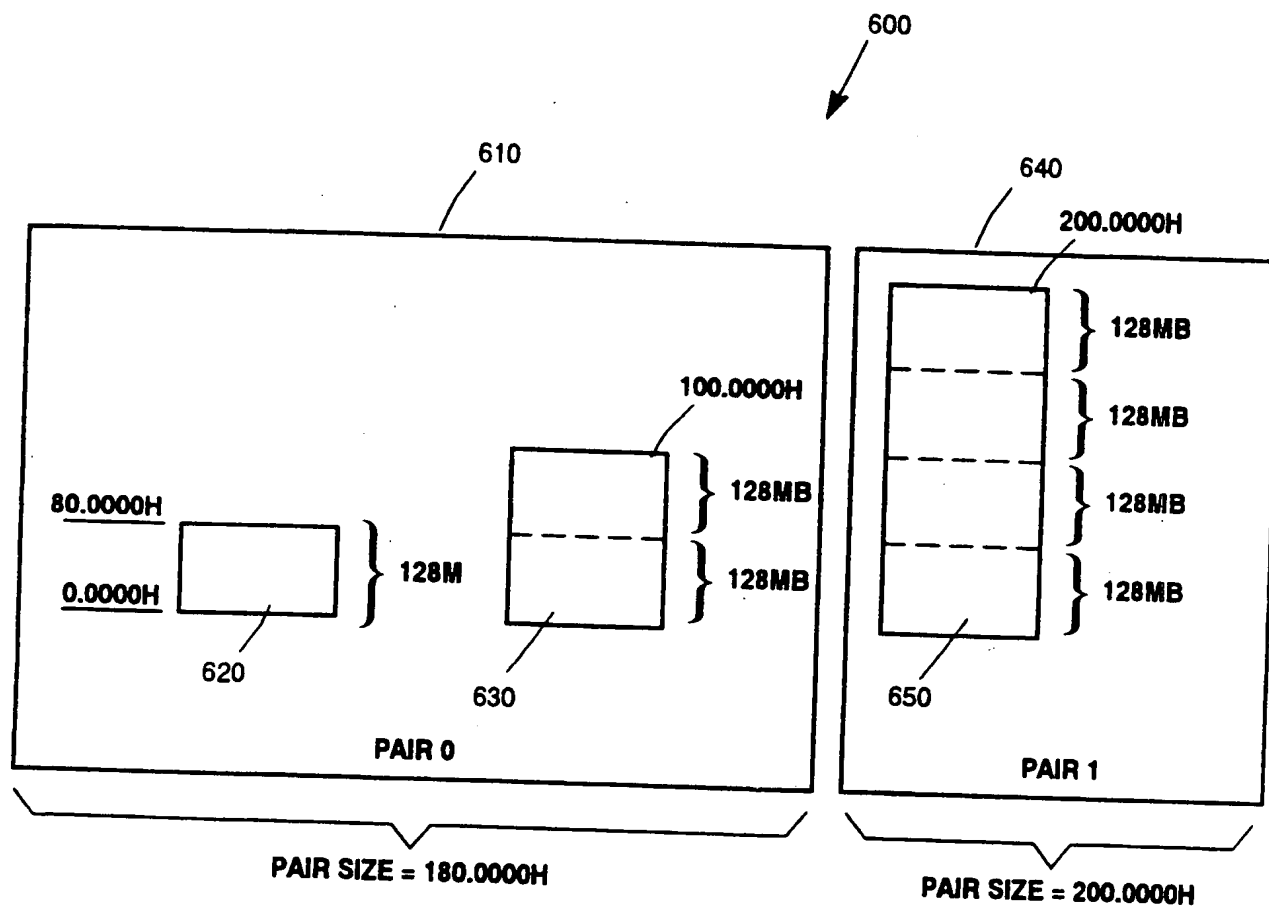


**Figure 4**

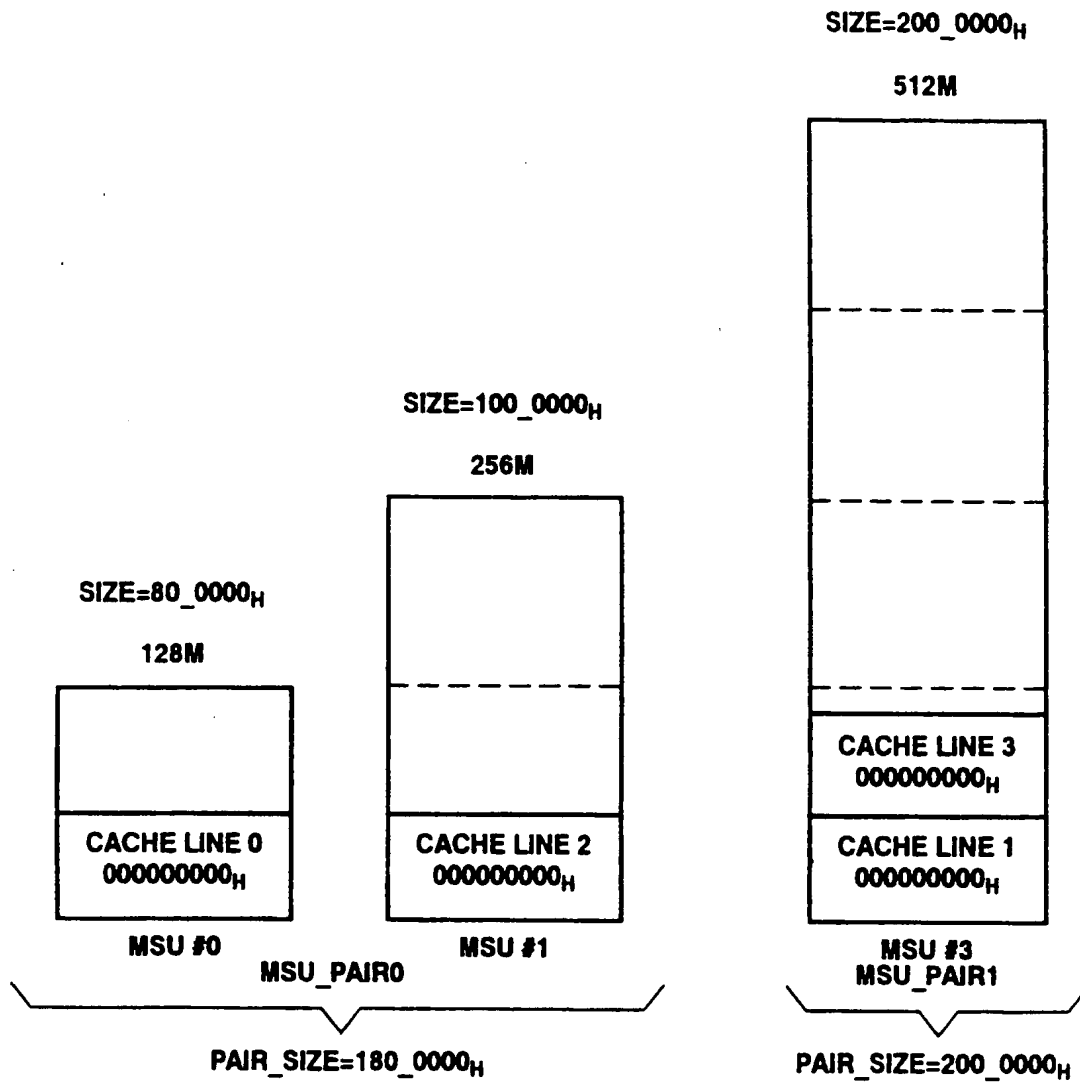
5/34



6/34

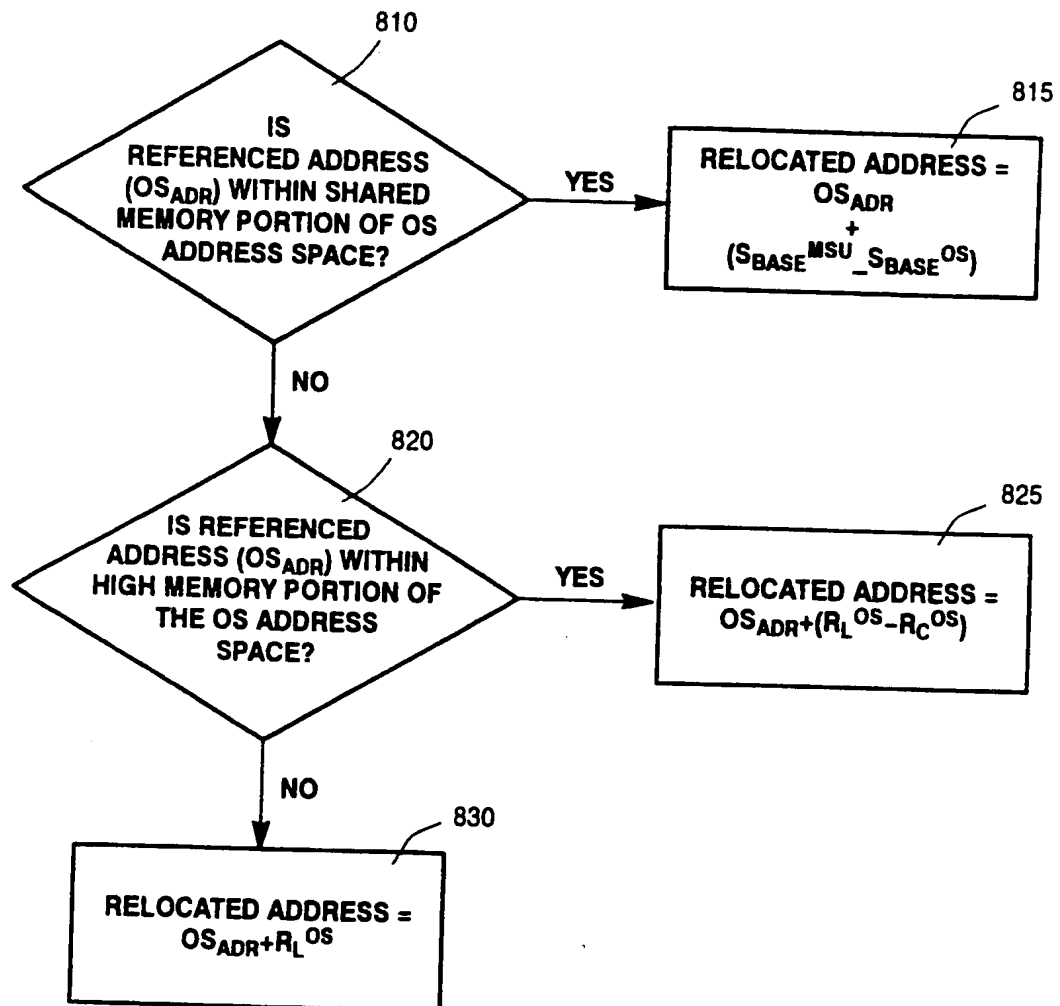
**Figure 6**

7/34

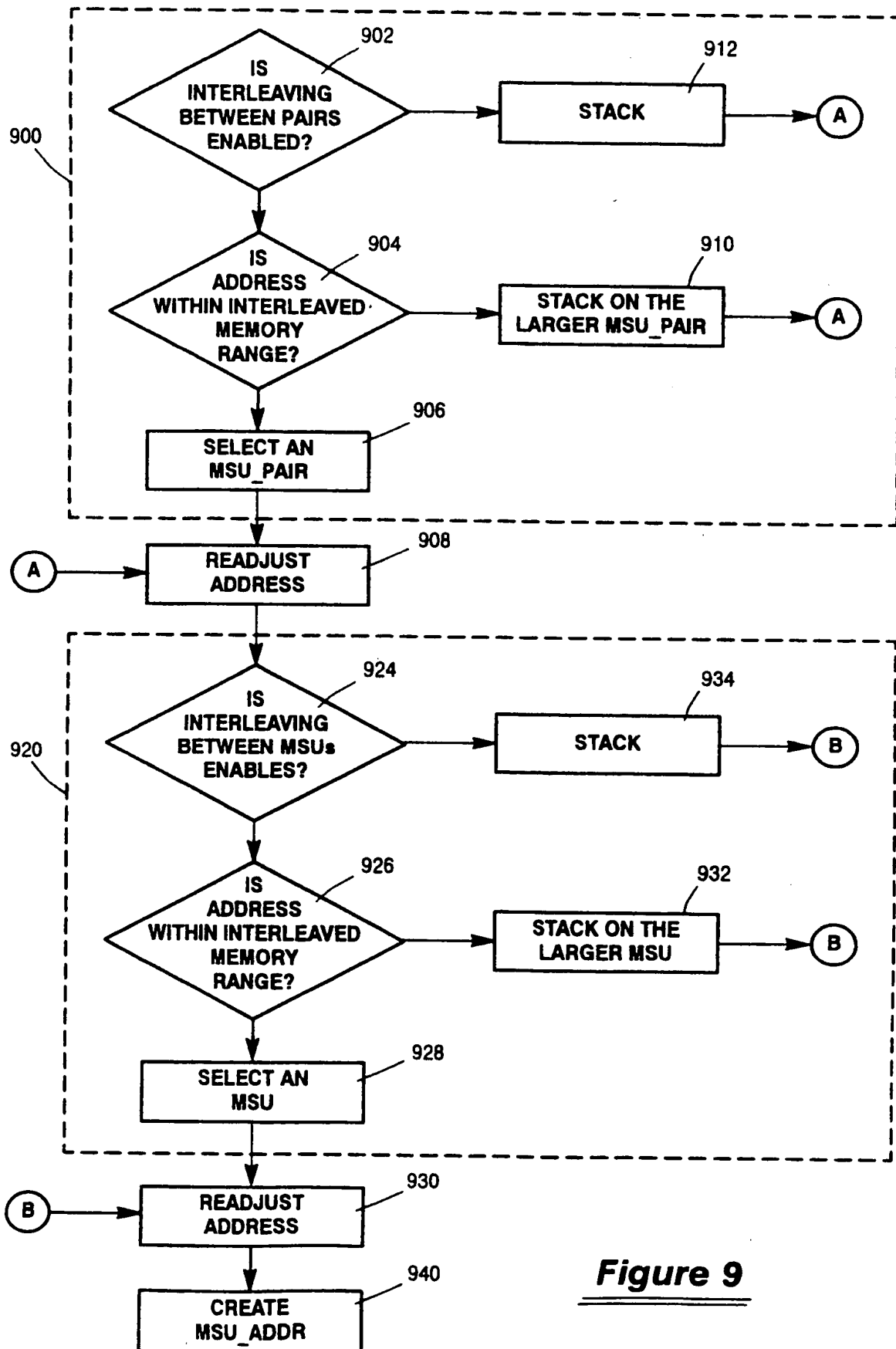


**Figure 7**

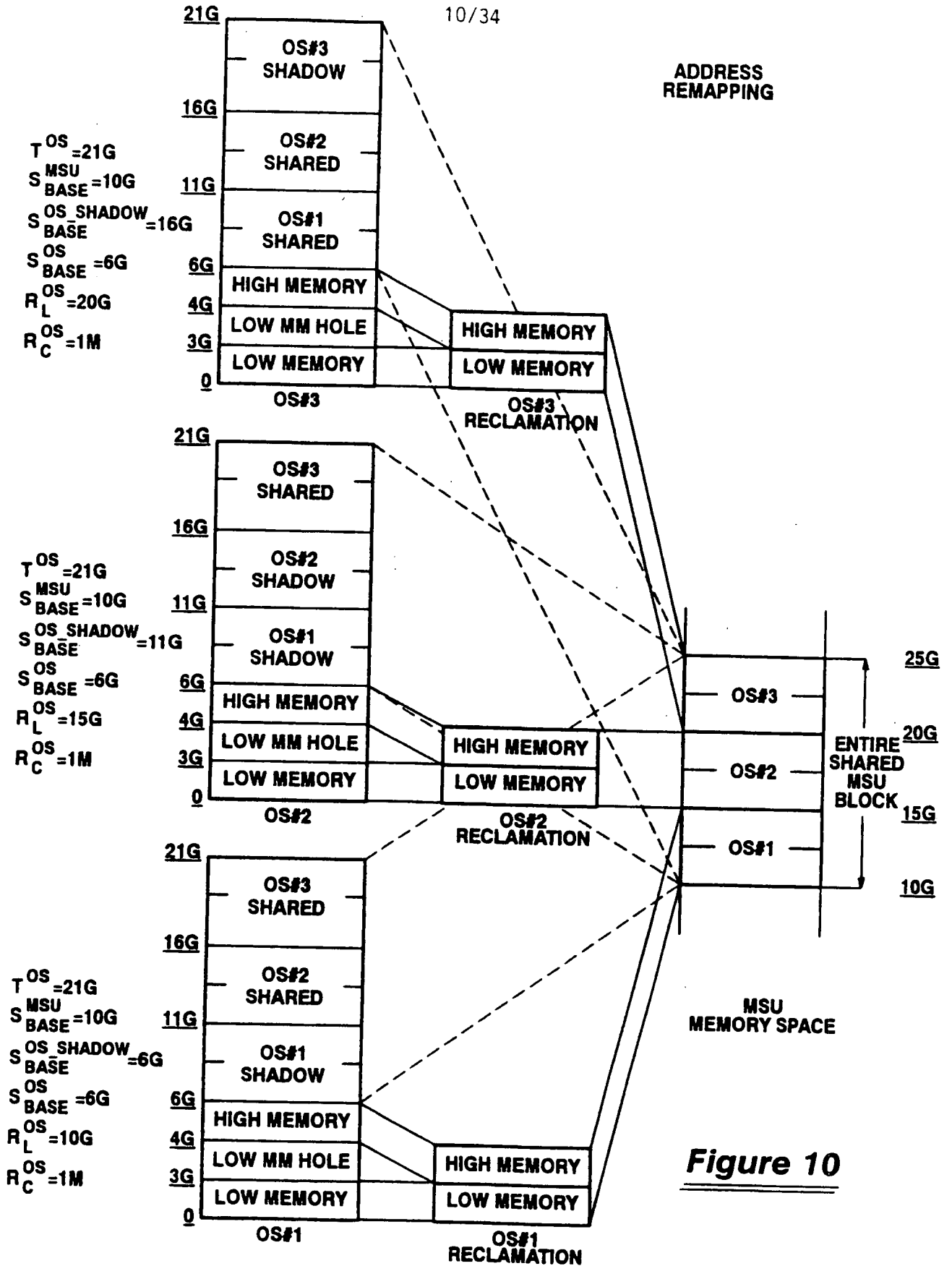
8/34

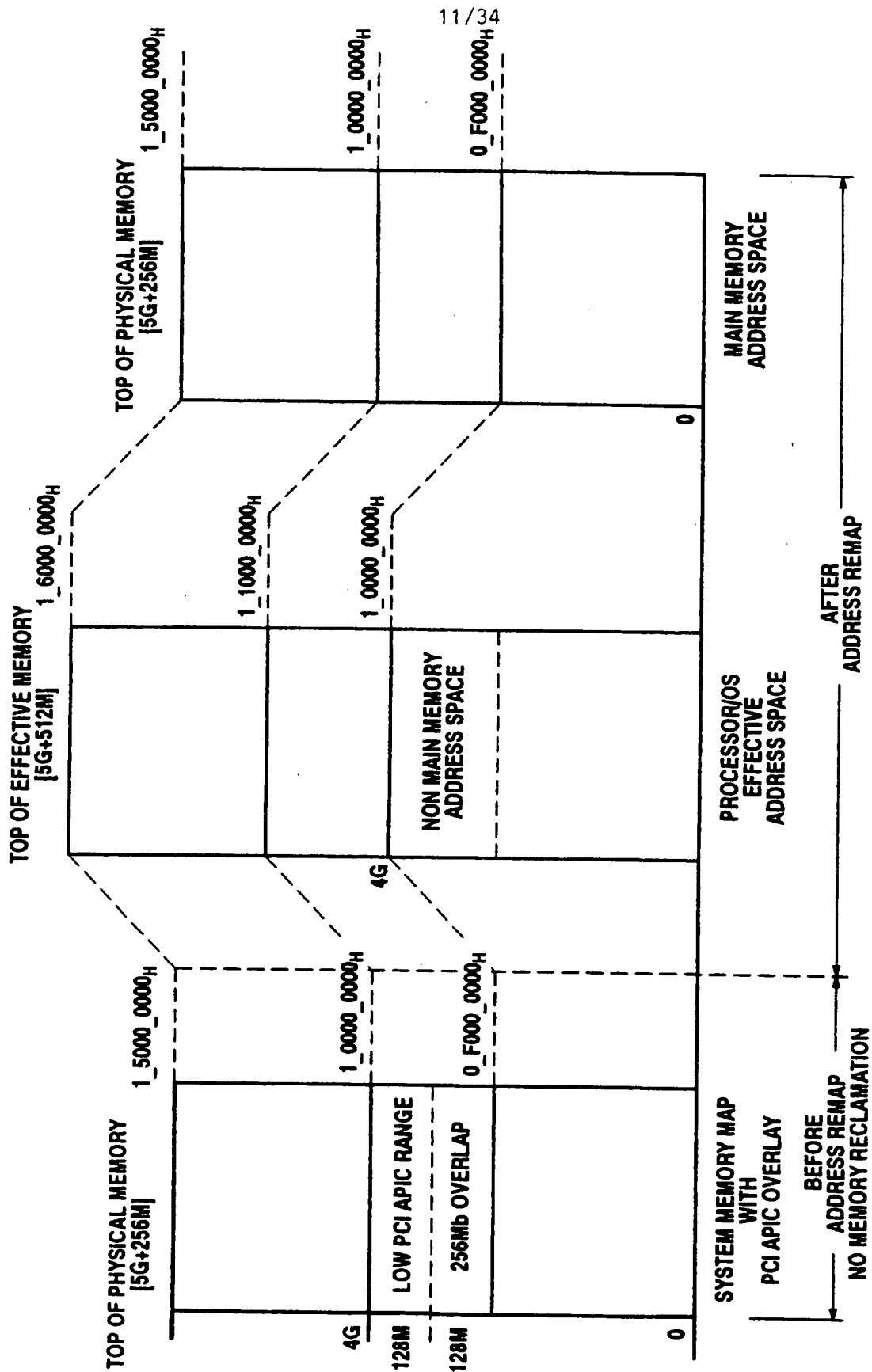
Figure 8

9/34

**Figure 9**







**Figure 11**

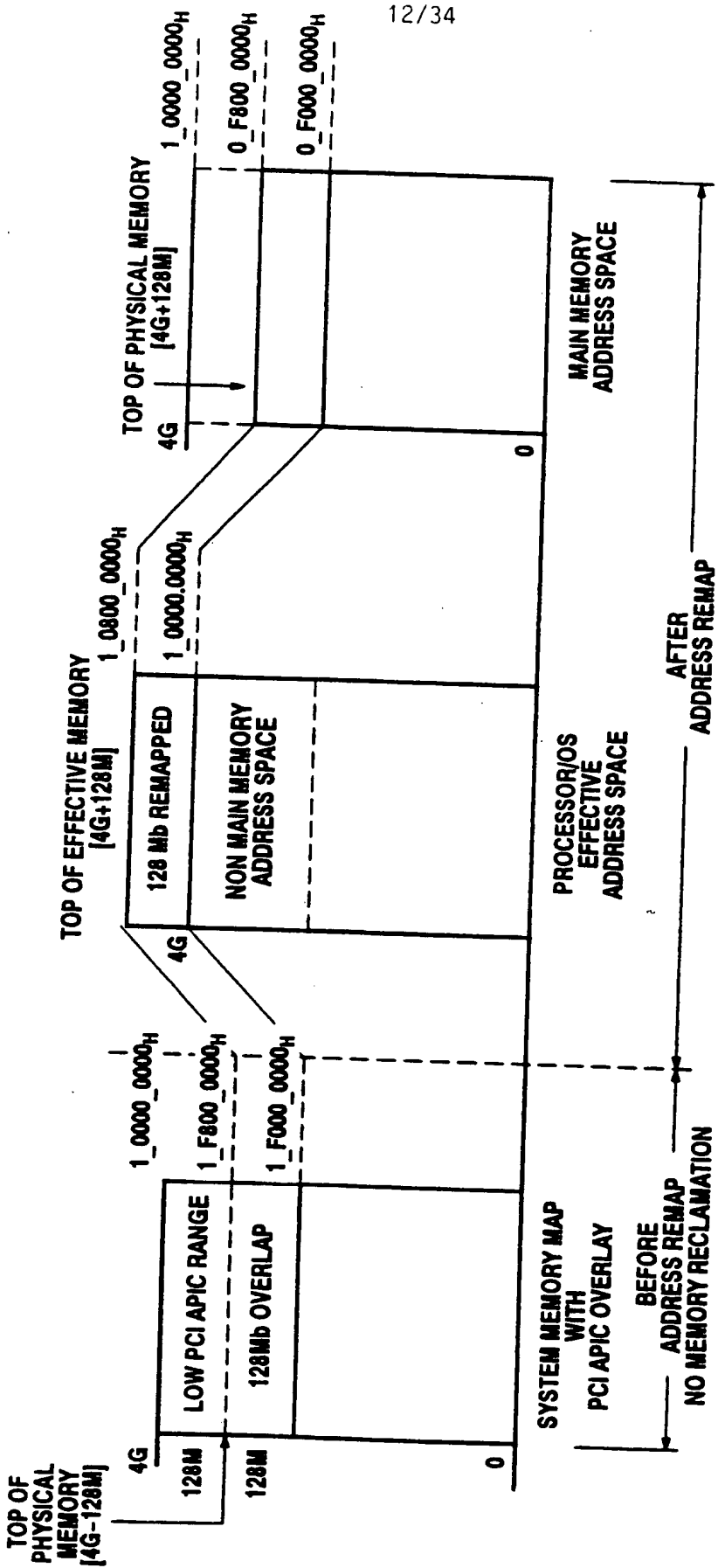
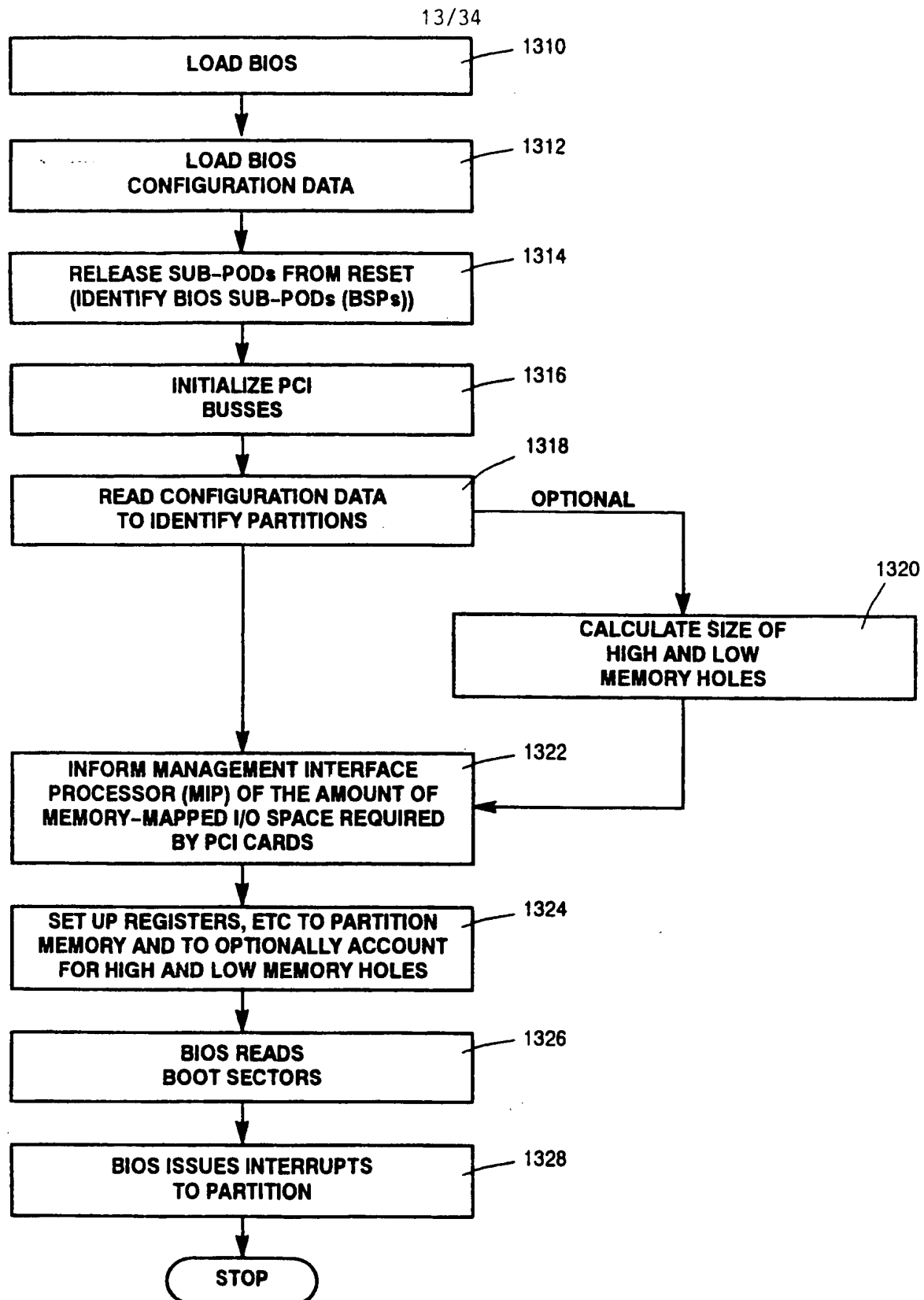
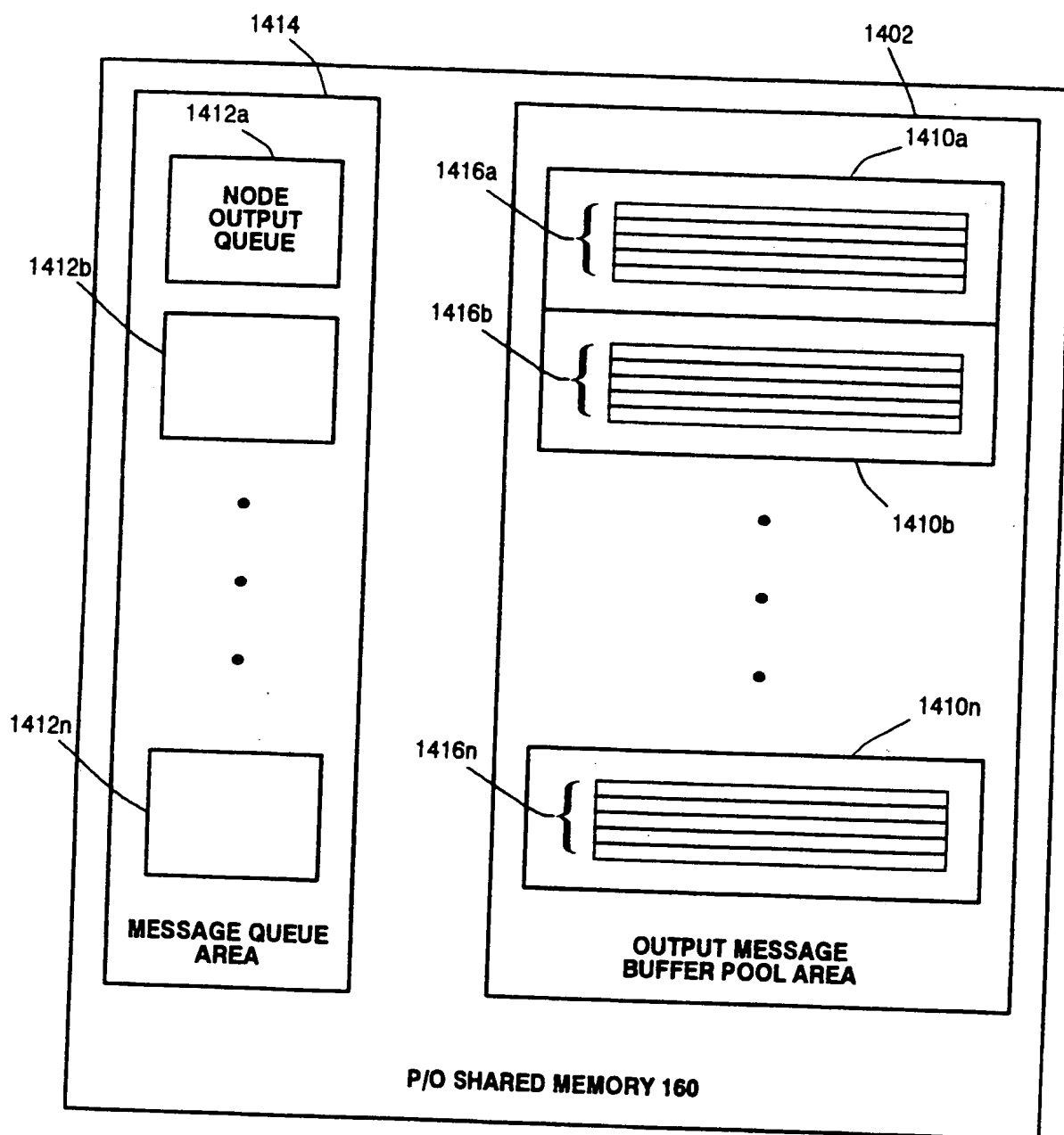


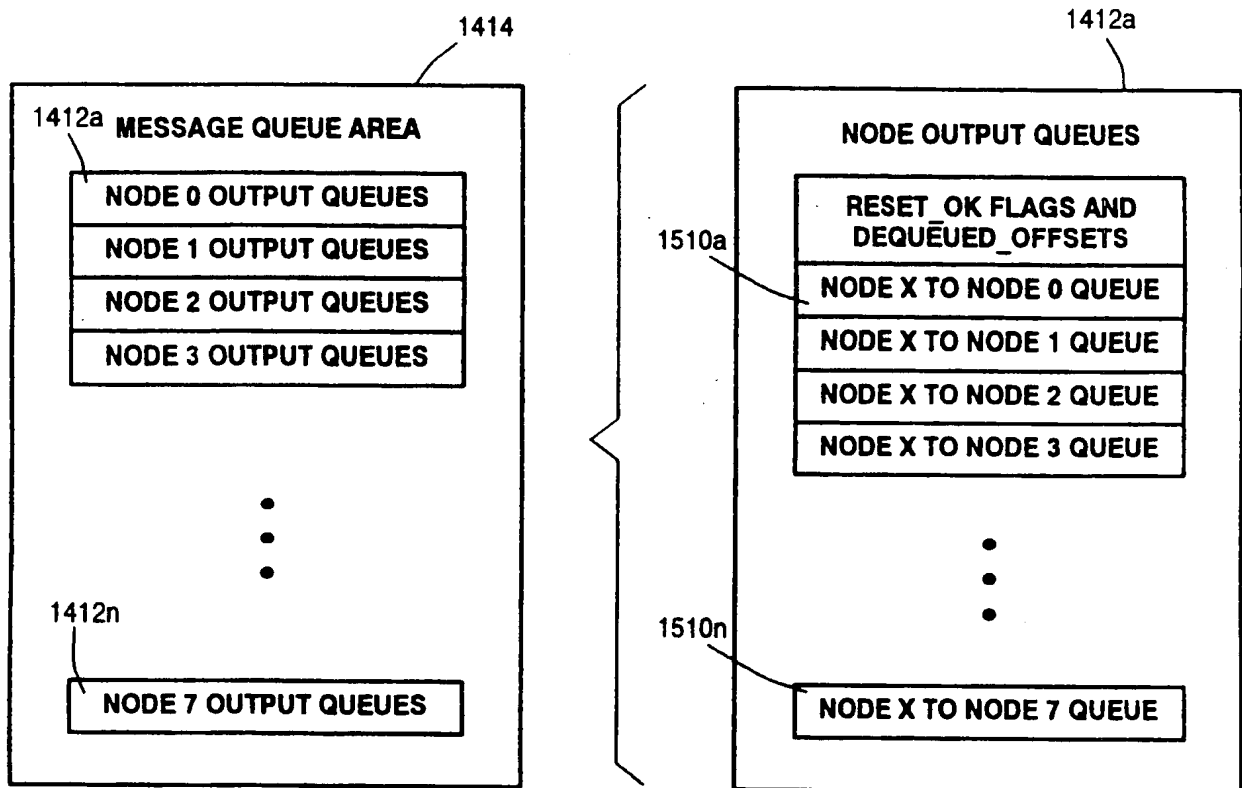
Figure 12

**Figure 13**

14/34

**Figure 14**

15/34

**Figure 15**

16/34

1412

A MORE DETAILED LOOK AT THE DEQUEUED OFFSETS  
AND THE MESSAGE QUEUES IS SHOWN BELOW:

	0	31	32 63				
0	RESERVED		NODE OS ID (EXAMPLES FOLLOW)				1610
			2 X	2 P N	S C I T	O M U N	1612
1-2	RESERVED		NODE MAC ADDRESS (12 HEX DIGITS WITH 2 DIGITS PER BYTE)				
3-7	RESERVED		RESERVED				
0	RESERVED		32 RESET_OK	39 RESERVED	40 RESERVED	47 48 63 DEQUEUED OFFSET FOR NODE 0	
1	RESERVED		RESET_OK	RESERVED	DEQUEUED_OFFSET FOR NODE 0		
2	RESERVED		RESET_OK	RESERVED	DEQUEUED_OFFSET FOR NODE 0		
				• • •			
7	RESERVED		RESET_OK	RESERVED	DEQUEUED_OFFSET FOR NODE 0		
START OF OUTPUT QUEUE TO NODE 0							
0	RESERVED		NEED_RESET	RESERVED	ENQUEUED_OFFSET FOR NODE 0		

P/O  
NODE-NODE  
QUEUE  
1510a

**Figure 16A**

17/34

1	RESERVED	MESSAGE BUFFER OFFSET			P/O NODE- TO-NODE QUEUE 1510a
2	RESERVED	MESSAGE BUFFER OFFSET			
		• • •			
511	RESERVED	MESSAGE BUFFER OFFSET			
START OF OUTPUT QUEUE TO NODE 1					1510
0	RESERVED	NEED_RESET	RESERVED	ENQUEUED OFFSET FOR NODE 1	
1	RESERVED	MESSAGE BUFFER OFFSET			
2	RESERVED	MESSAGE BUFFER OFFSET			
		• • •			
511	RESERVED	MESSAGE BUFFER OFFSET			
		• • •			
START OF OUTPUT QUEUE TO NODE 7					1510n
0	RESERVED	NEED_RESET	RESERVED	ENQUEUED OFFSET FOR NODE 7	
1	RESERVED	MESSAGE BUFFER OFFSET			
2	RESERVED	MESSAGE BUFFER OFFSET			
		• • •			
511	RESERVED	MESSAGE BUFFER OFFSET			

NODE\_OS\_ID IS A 4 CHARACTER STRING WITH ONE OF THE FOLLOWING VALUES:

- 'OS22' - OS2200 ARCHITECTURE
- 'MCP' - A-SERIES ARCHITECTURE
- 'UNIX' - INTEL ARCHITECTURE WITH A UNIX OPERATING SYSTEM
- 'NT' - INTEL ARCHITECTURE WITH MICROSOFT WINDOWS NT OPERATING SYSTEM

**Figure 16B**



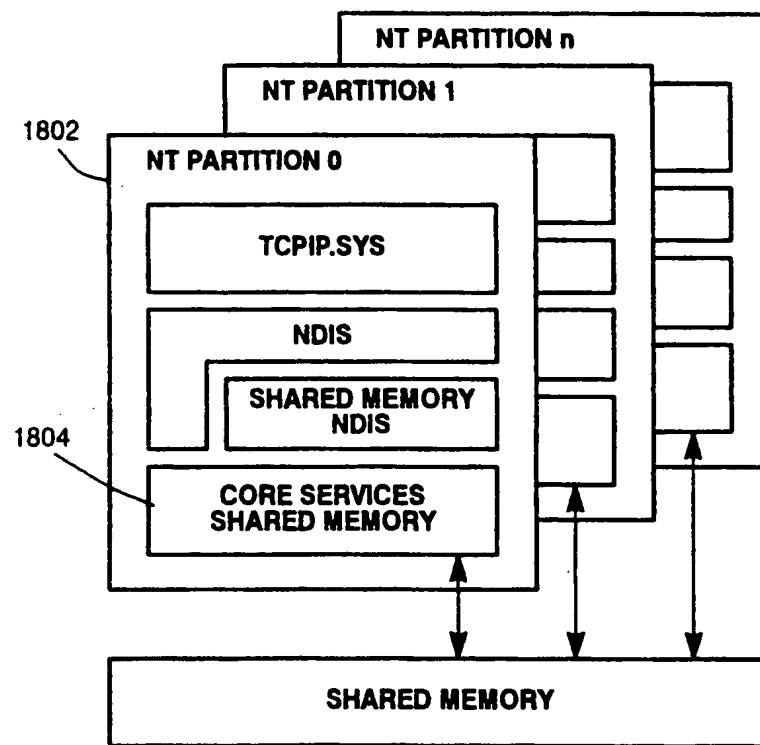
18/34

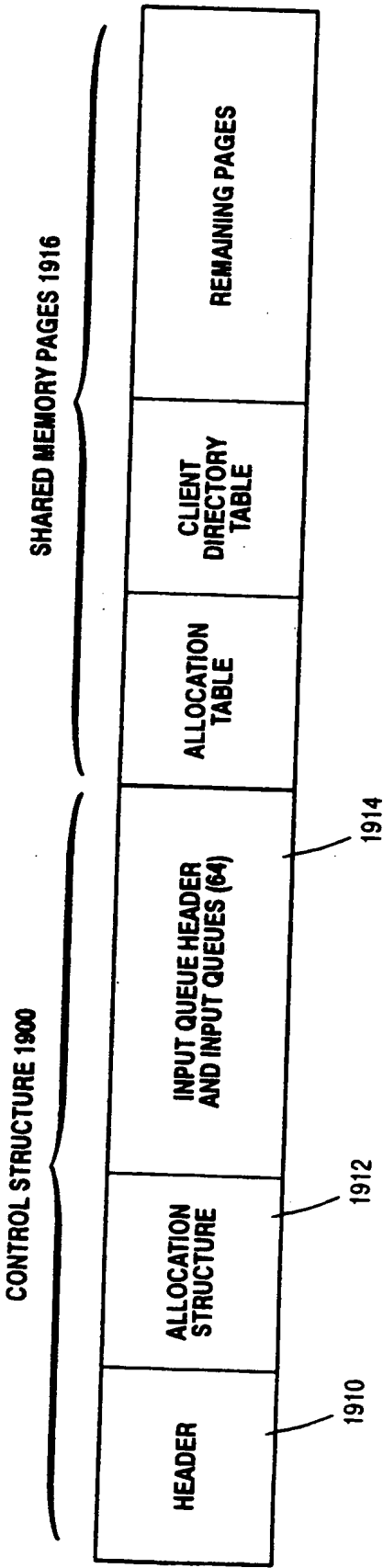
1416  
↓

0	0	RESERVED	31	32	WORD LENGTH OF BUFFER	63	1710
1		RESERVED			WORD LENGTH OF HEADER		
2		RESERVED		32	47	48	
				BYTE SKIP COUNT		63	
					BYTE TRANSFER COUNT		1712
3		RESERVED		BYTE SKIP COUNT		BYTE TRANSFER COUNT	
n		RESERVED		BYTE SKIP COUNT		BYTE TRANSFER COUNT	
m		RESERVED					
		RESERVED					1712
		RESERVED					
		RESERVED					
		RESERVED					
b-1		RESERVED					1712

Figure 17

19/34

**Figure 18**



**Figure 19**

21/34

1910

<b>CONTROL STRUCTURE HEADER CONTENTS</b>
<b>VERSION ID</b>
<b>SHARED MEMORY STATUS</b>
<b>PARTITION ID OF "MASTER PARTITION"</b>
<b>SHARED MEMORY PARTITION CHECK IN INTERVAL</b>
<b>CLIENT DIRECTORY TABLE HEADER</b>
<b>PARTITION INFORMATION (10 WORDS PER PARTITION)</b>

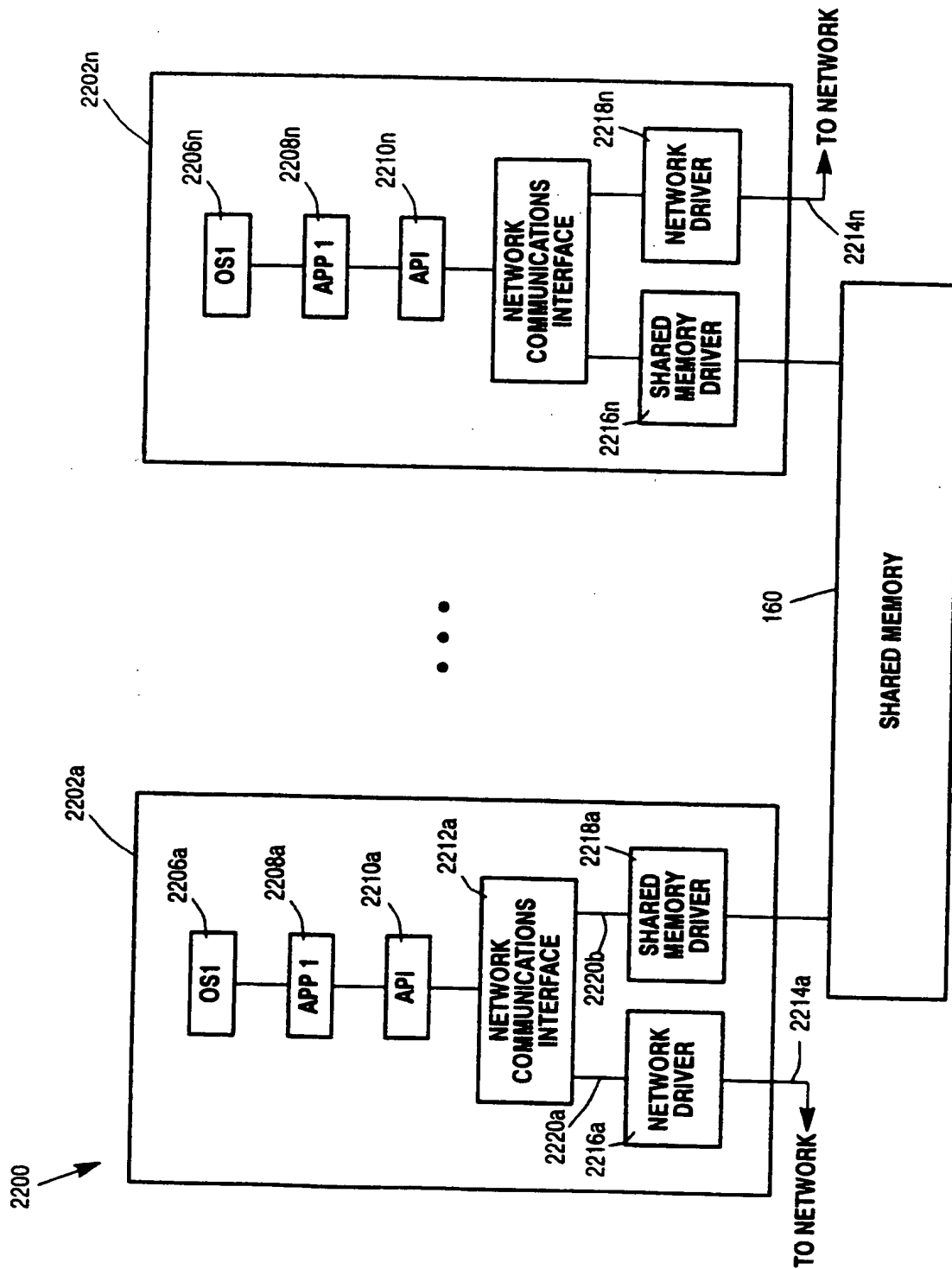
**Figure 20**

1912

<b>ALLOCATION STRUCTURE CONTENTS</b>
<b>ALLOCATION LOCK</b>
<b>LENGTH OF SHARED MEMORY AREA (IN 4K BYTES PAGES)</b>
<b>SHARED MEMORY PAGE POINTER</b>
<b>FREE PAGE LIST HEAD</b>
<b>ALLOCATION TABLE HEADER</b>

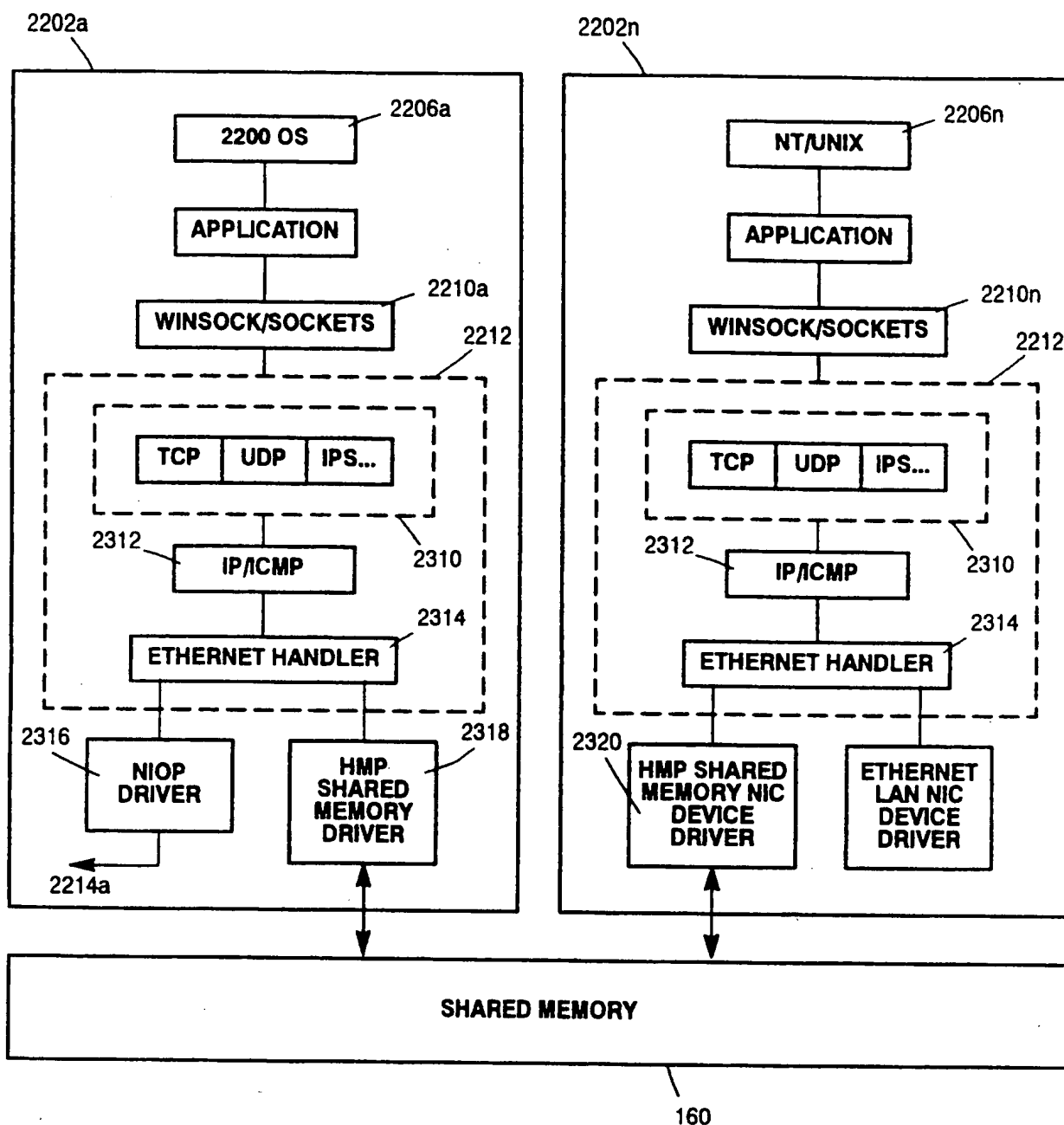
**Figure 21**

22/34

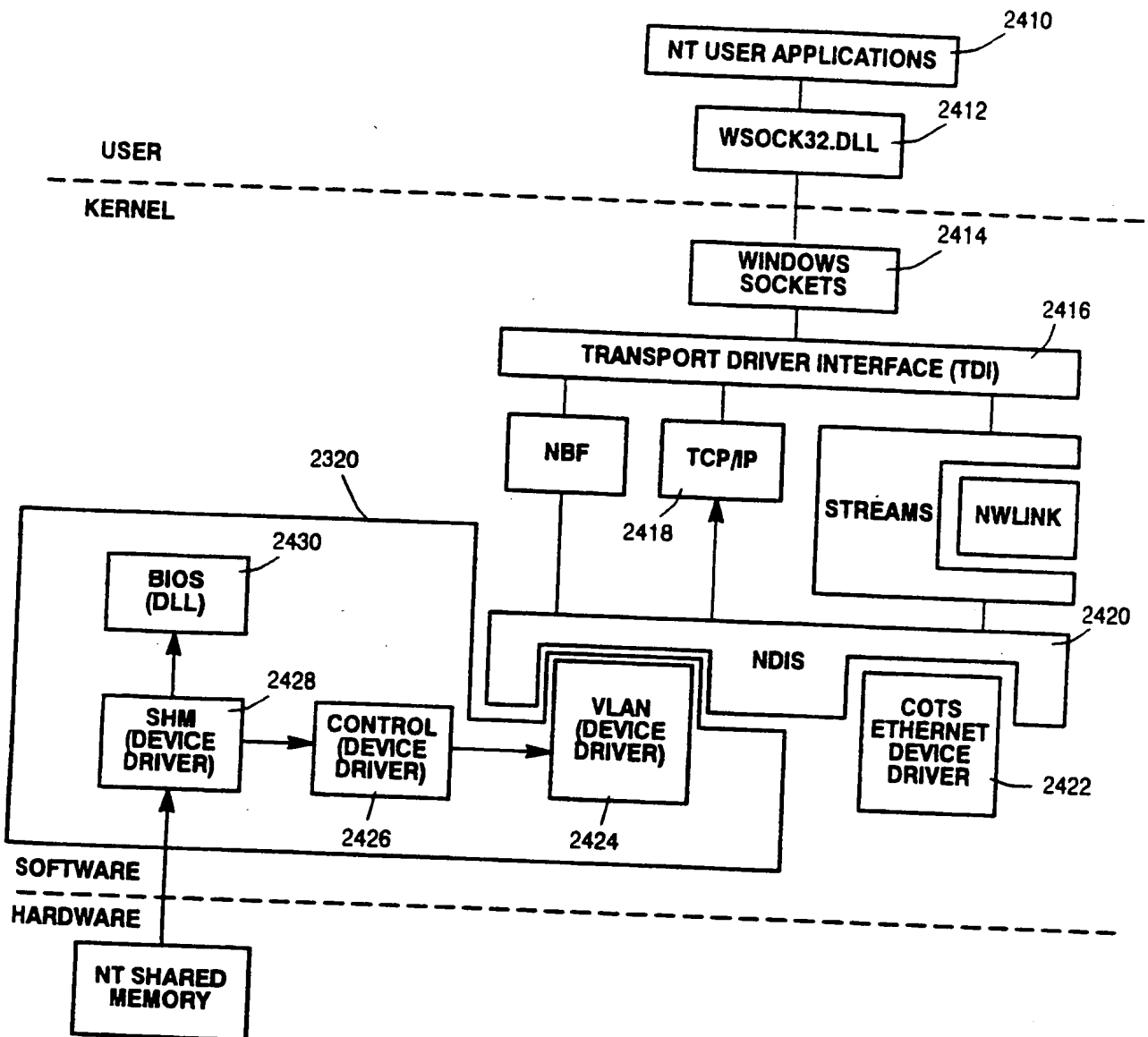


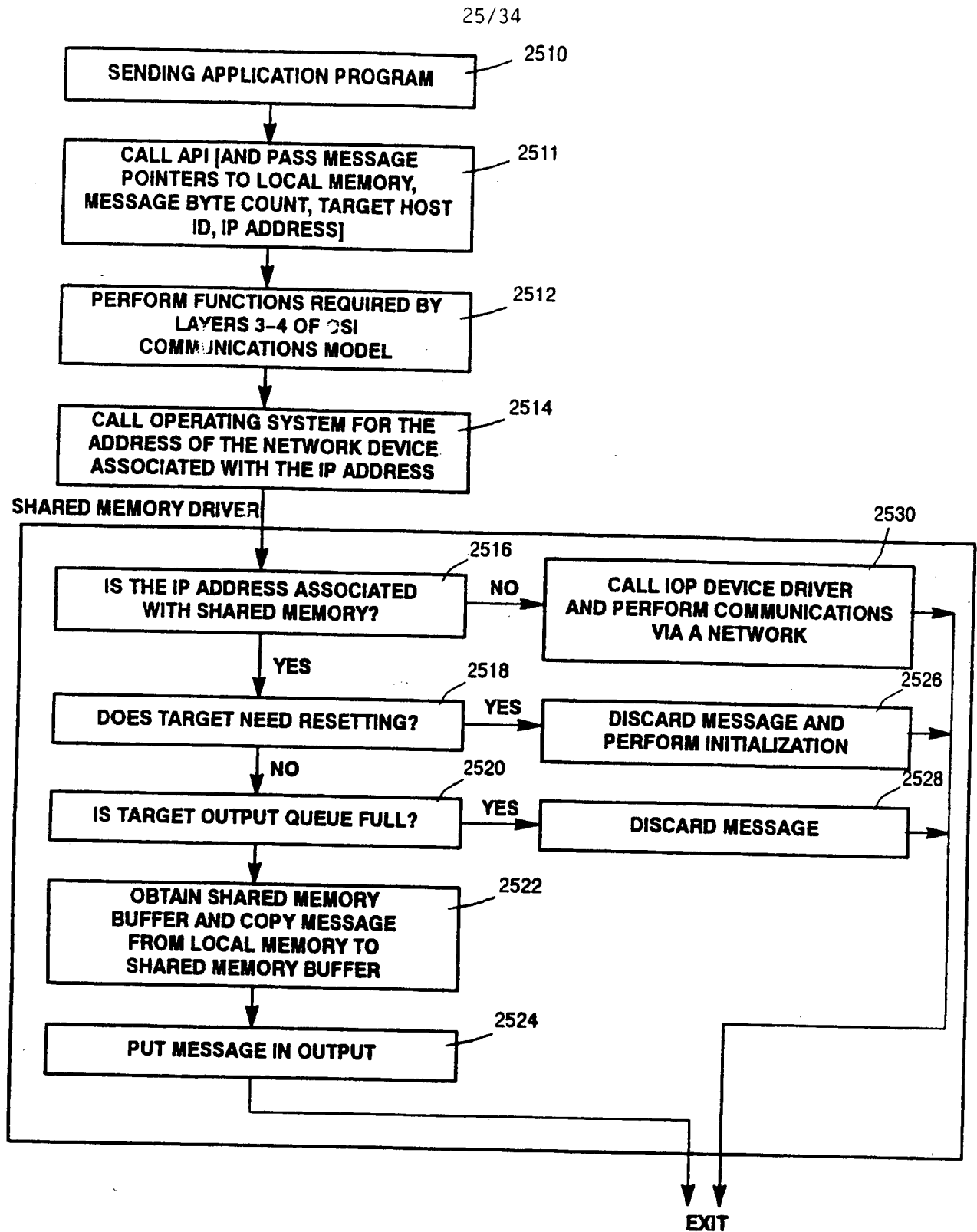
**Figure 22**

23/34

**Figure 23**

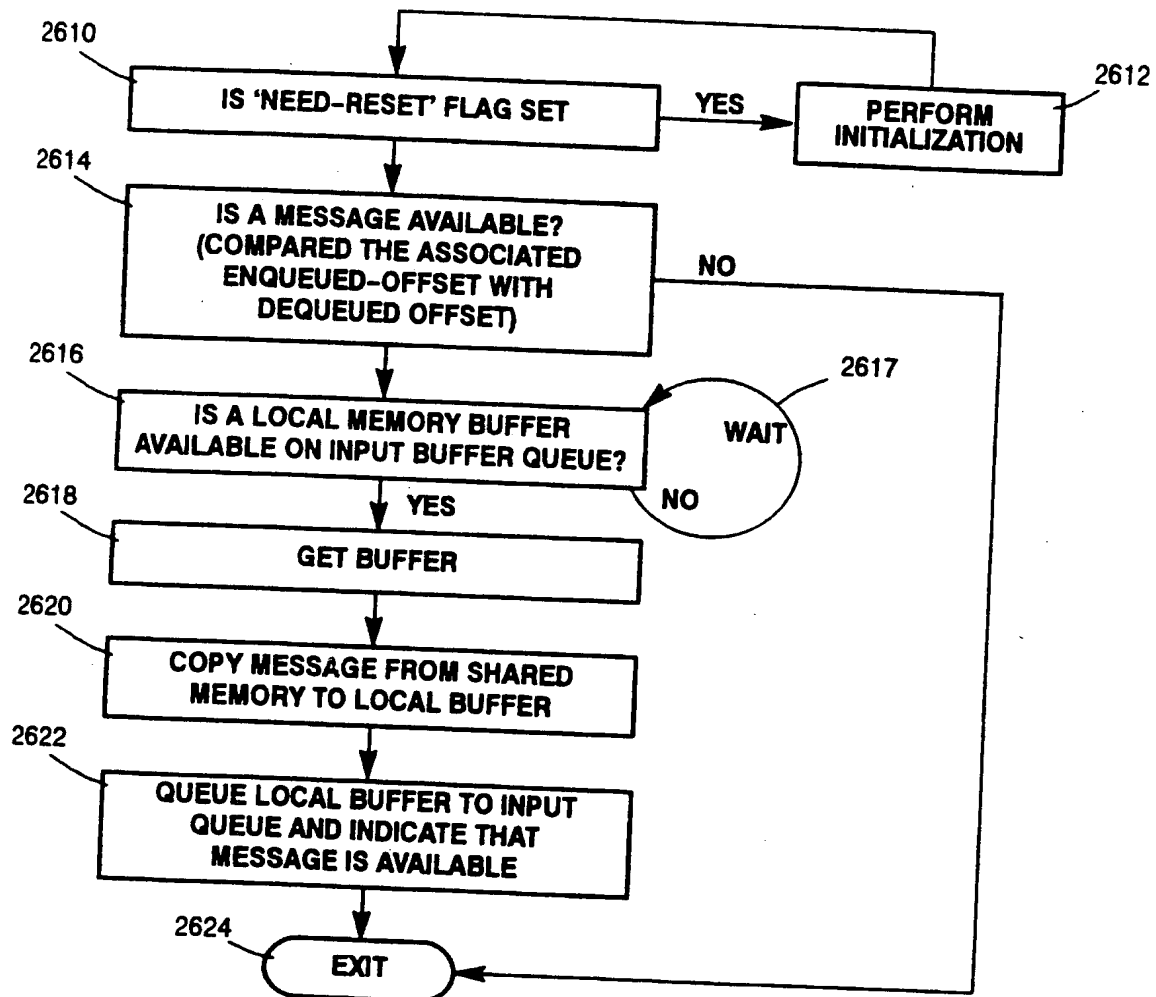
24/34

Figure 24

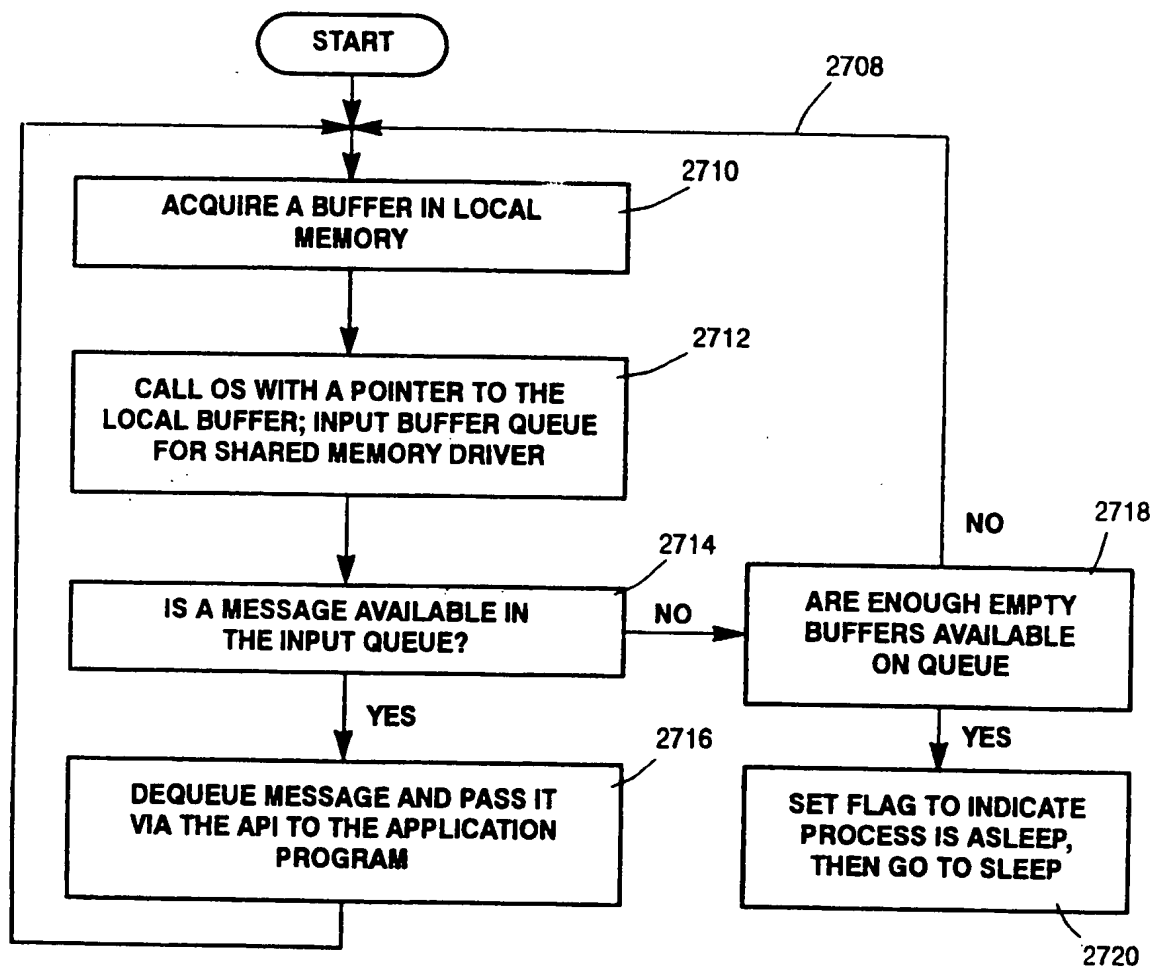
Figure 25



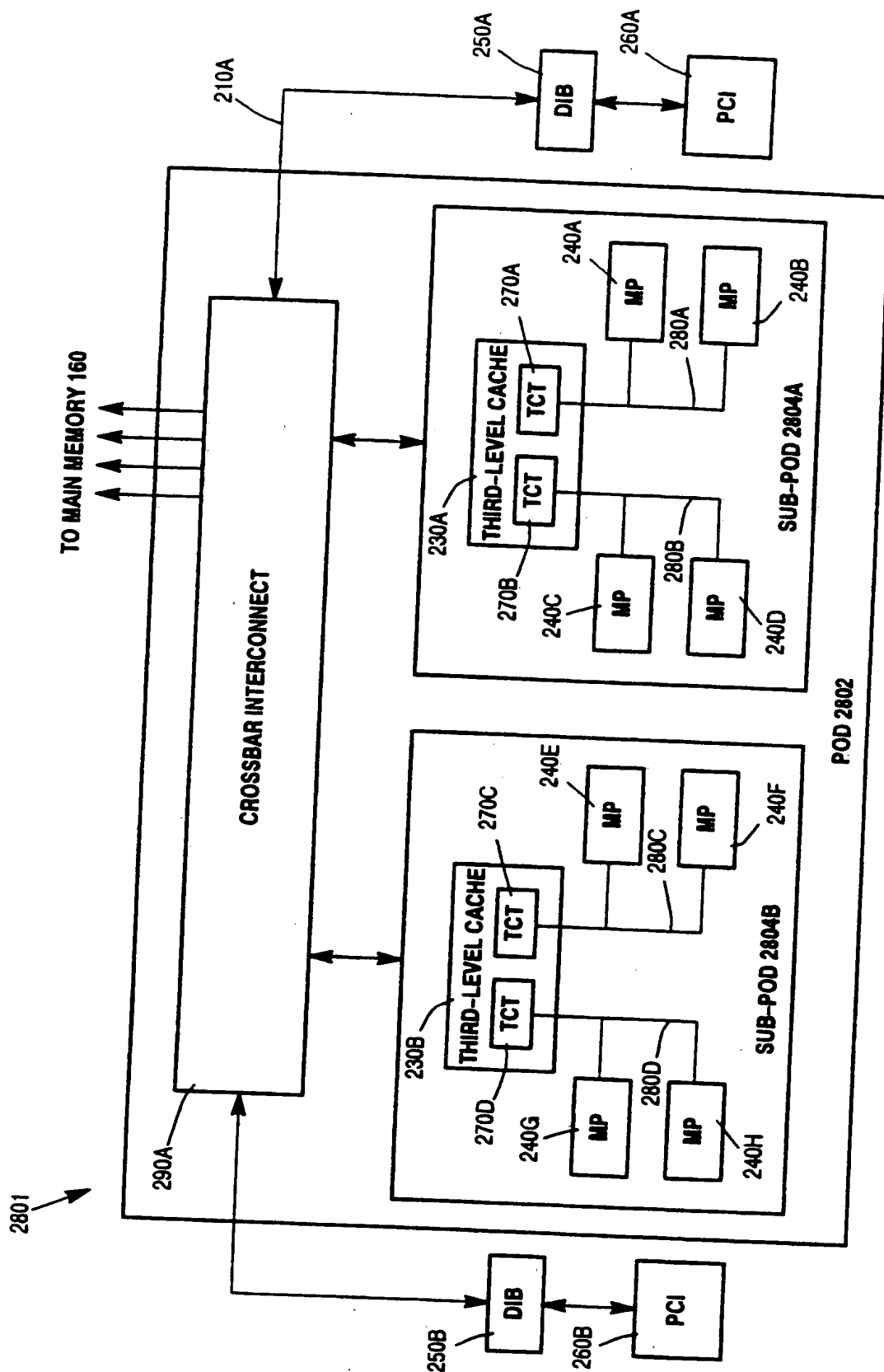
26/34

**Figure 26**

27/34

**Figure 27**

28/34

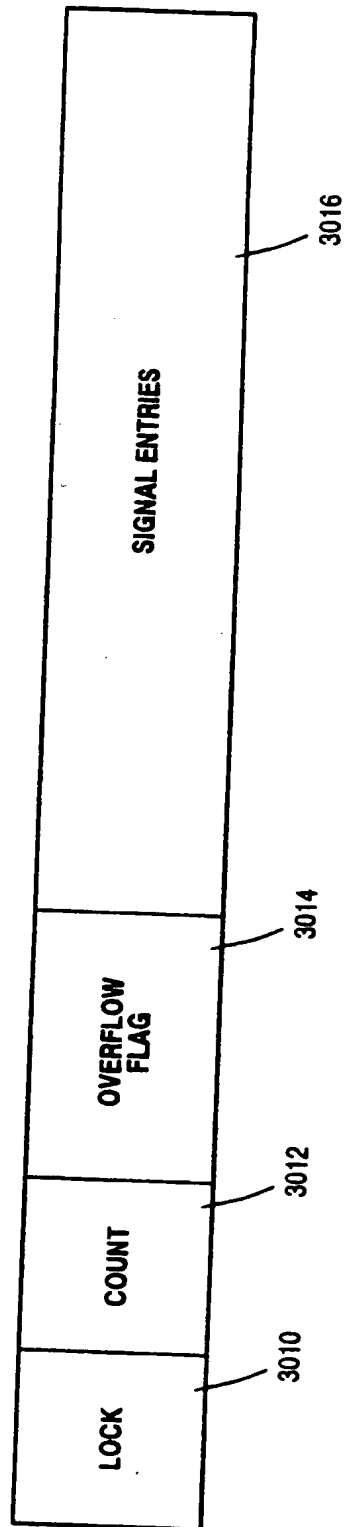


**Figure 28**

INPUT QUEUE HEADER CONTENTS
INPUT QUEUES POINTER
NUMBER OF INPUT QUEUES
INPUT QUEUE LENGTH
INPUT QUEUE SIGNAL SIZE
MAX NUMBER OF SIGNALS IN INPUT QUEUE

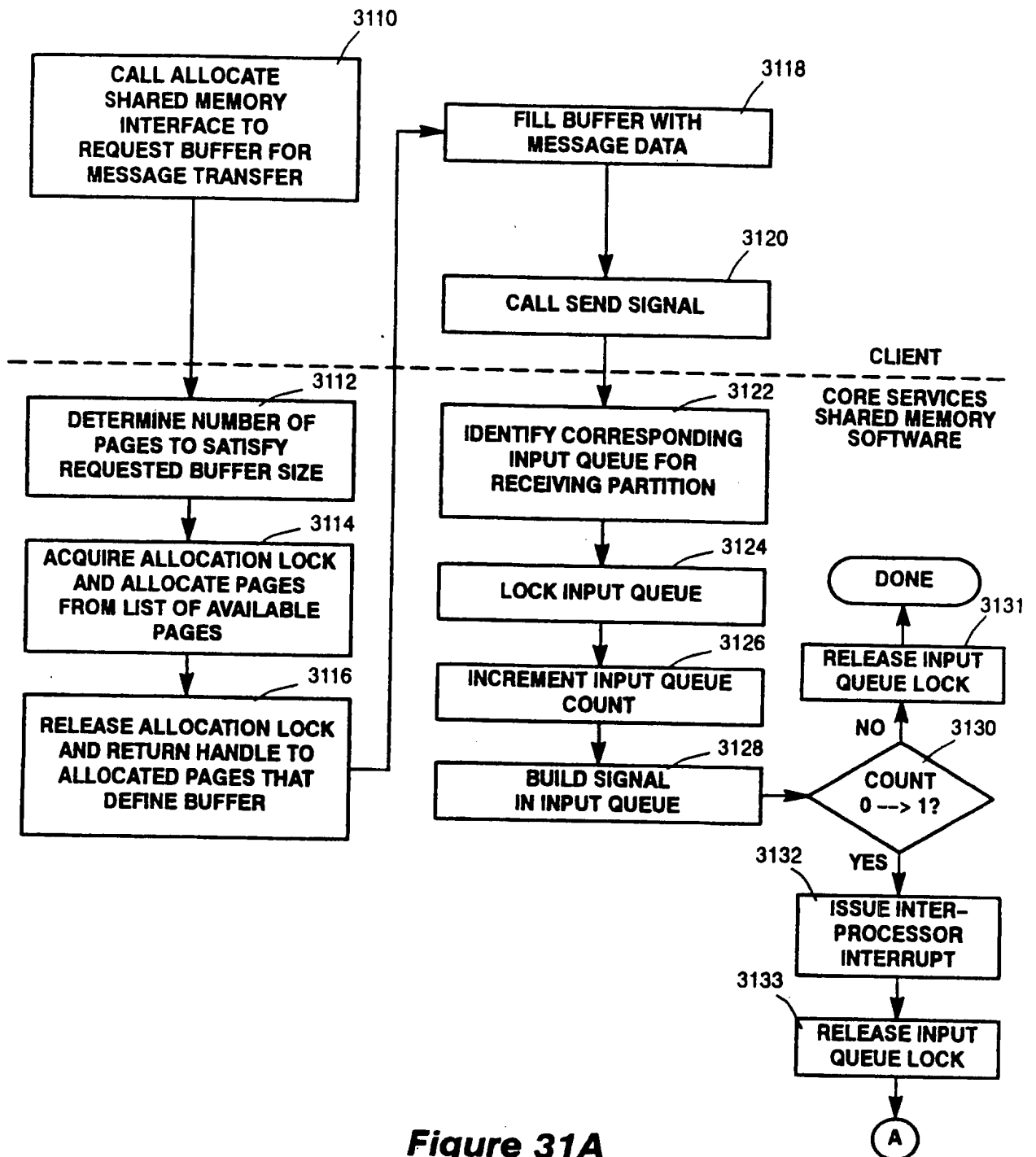
**Figure 29**

30/34

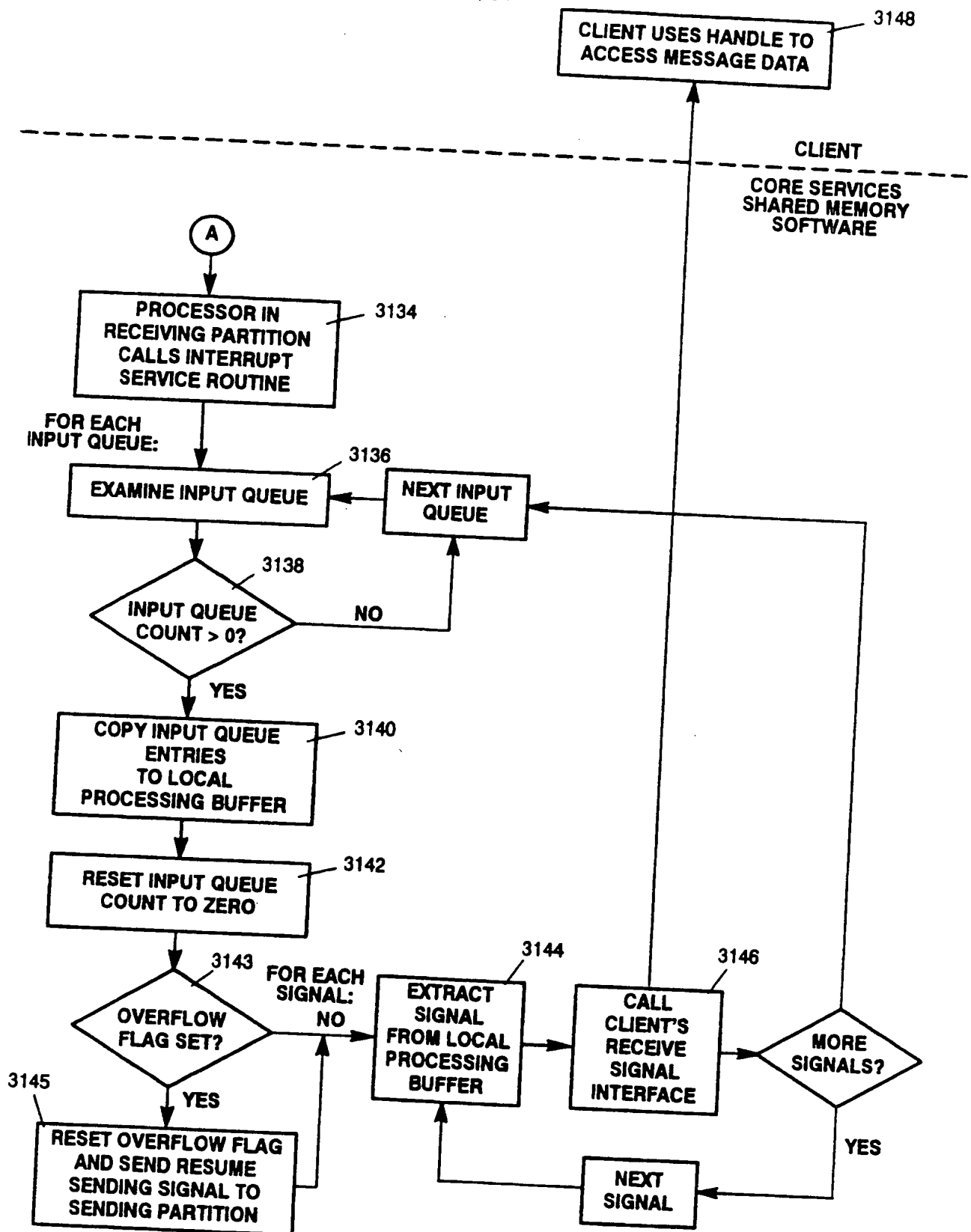


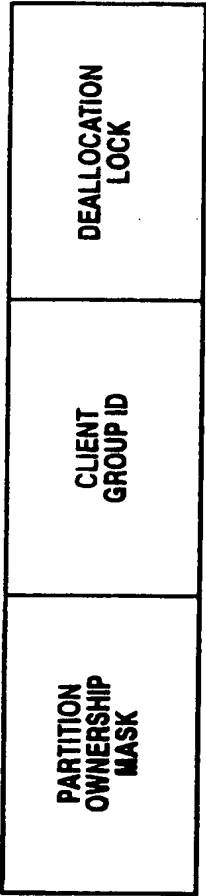
**Figure 30**

31/34

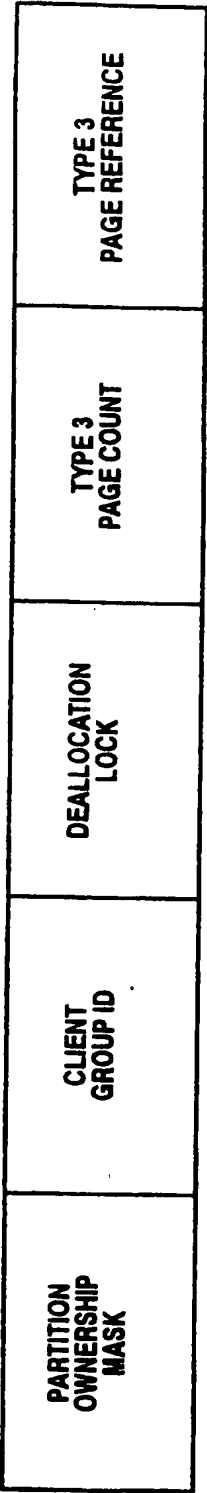


32/34

**Figure 31B**

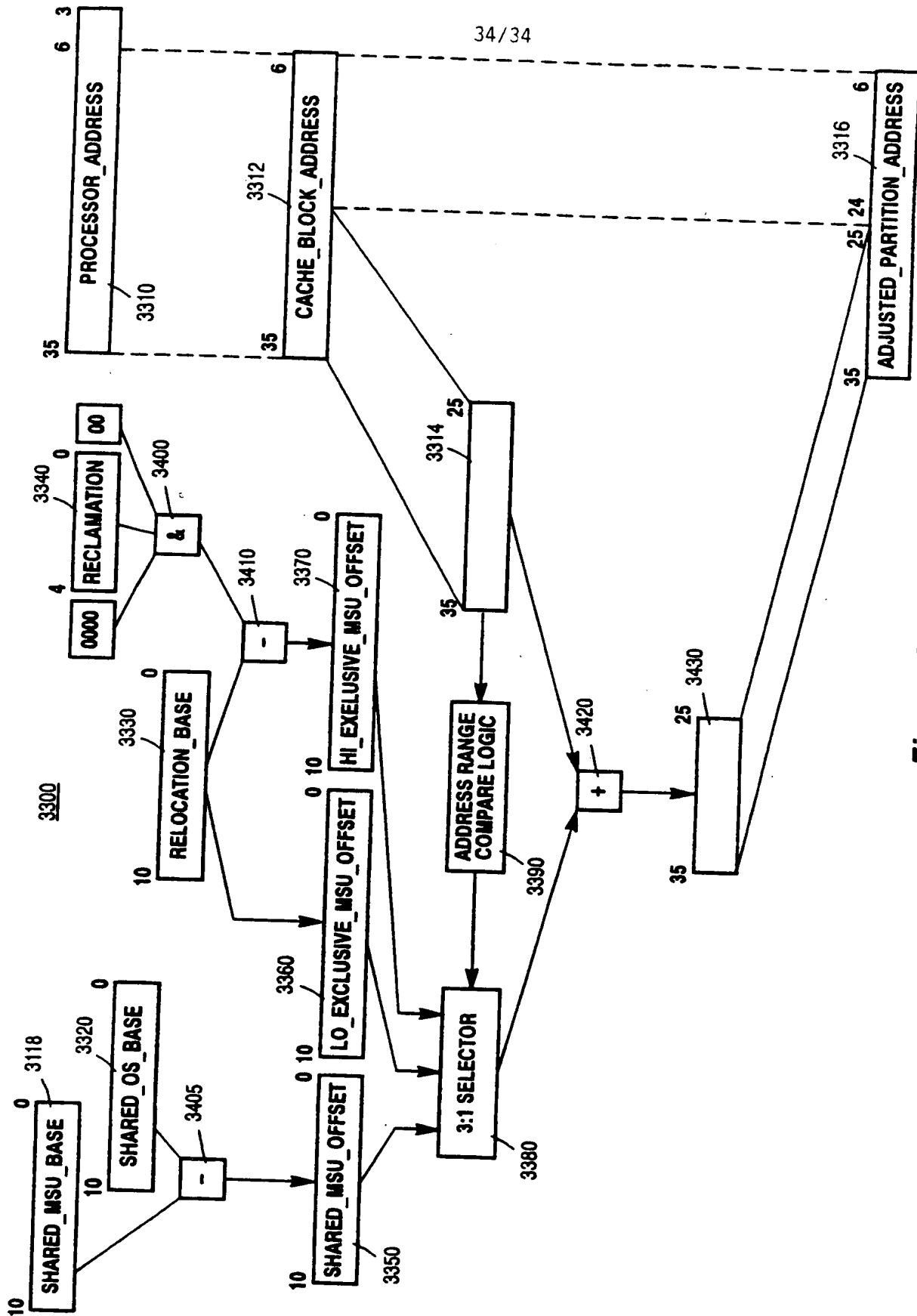


**Figure 32A**



**Figure 32B**



**Figure 33**

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
22 June 2000 (22.06.2000)

PCT

(10) International Publication Number  
**WO 00/36509 A3**

(51) International Patent Classification<sup>7</sup>: G06F 9/54,  
9/46, 12/02

(21) International Application Number: PCT/US99/30437

(22) International Filing Date:  
17 December 1999 (17.12.1999)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
09/215,424 18 December 1998 (18.12.1998) US

(71) Applicant: UNISYS CORPORATION [US/US]; Town-  
ship Line and Union Meeting Roads, P.O. Box 500, Blue  
Bell, PA 19424-0001 (US).

(72) Inventors: GULICK, Robert, C.; 331 Fairview Road,  
Glenmoore, PA 19343 (US). MORRISSEY, Douglas, E.;  
3923 Short Hill Drive, Allentown, PA 18104 (US). CAL-  
DARALE, Charles, Raymond; 2843 Brighton Avenue  
NE, Minneapolis, MN 55418 (US). VESSEY, Bruce,  
Alan; 916 Noble Drive, Downingtown, PA 19335 (US).

RUSS, Craig, F.; 30 Robins Lane, Berwyn, PA 19312  
(US). TROXELL, Eugene, W.; 273 Candlebrook Road,  
King of Prussia, PA 19406-1808 (US). MIKKELSEN,  
Hans, Christian; 14661 Afton Blvd. South, Afton, MN  
55001 (US). MAUER, Sharon, M.; 1064 Westwood  
Drive, West Chester, PA 19382 (US). CONNELL, Mau-  
reen, P.; 2124 Hemlock Road, Norristown, PA 19403  
(US). HUNTER, James, R.; 461 East Radnor Ct., Down-  
ingtown, PA 19335-1734 (US).

(74) Agents: STARR, Mark, T. et al.; Unisys Corporation,  
Township Line and Union Meeting Roads, P.O. Box 500,  
Blue Bell, PA 19424-0001 (US).

(81) Designated States (*national*): BR, CA, JP.

(84) Designated States (*regional*): European patent (AT, BE,  
CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC,  
NL, PT, SE).

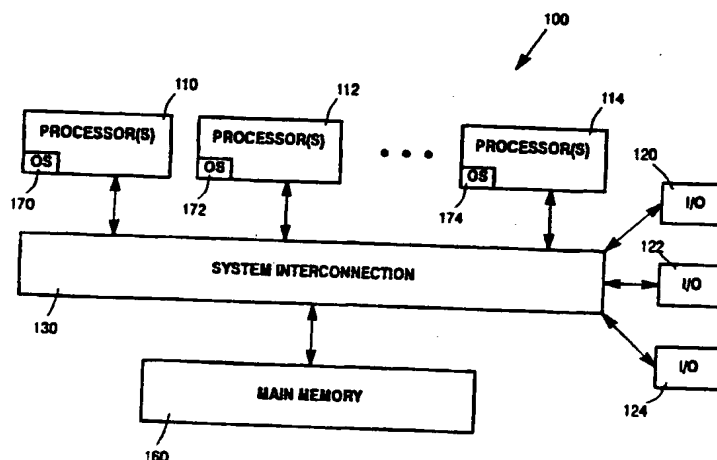
Published:

— With international search report.

(88) Date of publication of the international search report:  
19 April 2001

[Continued on next page]

(54) Title: COMPUTER SYSTEM AND METHOD FOR OPERATING MULTIPLE OPERATING SYSTEMS IN DIFFERENT PARTITIONS OF THE COMPUTER SYSTEM AND FOR ALLOWING THE DIFFERENT PARTITIONS TO COMMUNICATE WITH ONE ANOTHER THROUGH SHARED MEMORY



(57) Abstract: A computer system comprises a plurality of processing modules that can be configured into different partitions within the computer system, and a main memory. Each partition operates under the control of a separate operating system. At least one shared memory window is defined within the main memory to which multiple partitions have shared access, and each partition may also be assigned an exclusive memory window. Program code executing on different partitions enables those partitions to communicate with each other through the shared memory window. Means are also provided for mapping the physical address space of the processors in each partition to the respective exclusive memory windows assigned to each partition, so that the exclusive memory windows assigned to each partition appear to the respective operating systems executing on those partitions as if they all start at the same base address.

WO 00/36509 A3



*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

## INTERNATIONAL SEARCH REPORT

Int Application No  
PCT/US 99/30437

## A. CLASSIFICATION OF SUBJECT MATTER

IPC 7 G06F9/54 G06F9/46 G06F12/02

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, IBM-TDB, INSPEC

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y  A	US 5 123 094 A (MACDOUGALL MYRON H) 16 June 1992 (1992-06-16) column 8, line 18 - line 56  column 9, line 27 - column 10, line 8; figures 6,7  ----- -/-	1,5-7, 11,34-36 8-10,12, 19,20, 22,37, 44,45, 48,76, 77,83,84 3,4,12, 37,38,79

☒ Further documents are listed in the continuation of box C.☒ Patent family members are listed in annex.

## \* Special categories of cited documents:

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

- \*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- \*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- \*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- \*&\* document member of the same patent family

Date of the actual completion of the international search

14 November 2000

Date of mailing of the international search report

08 12. 2000

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Bijn, K

## INTERNATIONAL SEARCH REPORT

PCT/US 99/30437

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 123 101 A (SINDHU PRADEEP S) 16 June 1992 (1992-06-16)	1,2,56, 58,62, 66,70, 73,91,95
Y	column 6, line 16 -column 7, line 44; figure 1	48, 52-55, 74-77, 83,84, 86-89
A		34,57, 59,67, 72,90,92
X	US 5 652 885 A (REED DENNIS F ET AL) 29 July 1997 (1997-07-29)	1,4-6,34
Y	column 4, line 62 -column 6, line 23	8-10,20, 22,75
A	column 9, line 4 -column 12, line 9	2,7-11, 35,36, 73,78, 80-82
X	US 5 276 896 A (RIMMER TODD M ET AL) 4 January 1994 (1994-01-04)	1,5, 26-29
Y	column 3, line 62 -column 7, line 5	52-55, 86-89
A	column 9, line 58 -column 11, line 23	3,4,30, 31
Y	US 5 142 683 A (GERBEHY JAY L ET AL) 25 August 1992 (1992-08-25)	12,37
A	abstract column 9, line 65 -column 11, line 36	41,45
X	US 5 689 713 A (NORMOYLE KEVIN B ET AL) 18 November 1997 (1997-11-18)	1,5,6,34
Y	column 7, line 45 -column 12, line 55	19,44,45 3,4,7, 11, 13-16, 19,35,85
X	EP 0 444 376 A (IBM) 4 September 1991 (1991-09-04)	1,3,5,26
Y	column 3, line 45 -column 4, line 19	74
A	column 7, line 18 - line 54 column 9, line 33 -column 11, line 44	27-33
A	US 5 233 701 A (NAKATA AKIHIKO) 3 August 1993 (1993-08-03)	21
	column 2, line 28 -column 3, line 10 -/-	

# INTERNATIONAL SEARCH REPORT

Int. Application No  
PCT/US 99/30437

## C. (Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 590 301 A (GUENTHNER RUSSELL W ET AL) 31 December 1996 (1996-12-31)	56-59, 62, 64-67, 70,72, 73,
Y	column 2, line 13 -column 7, line 17	90-92,95 60,61, 63,68, 69,71, 93,94,96
Y	US 5 117 350 A (PEIFFER JR ROBERT E ET AL) 26 May 1992 (1992-05-26) column 9, line 15 - line 43 column 12, line 61 -column 14, line 2; figure 7	63,71,96
Y	US 5 687 342 A (KASS WILLIAM J) 11 November 1997 (1997-11-11) column 2, line 18 -column 4, last line	60,61, 68,69, 93,94

# INTERNATIONAL SEARCH REPORT

national application No.  
PCT/US 99/30437

## Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:  
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:  
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:
3. ☐ Claims Nos.:  
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

## Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☒ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest.
- ☒ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. Claims: 1-55 73-89

Communication via shared memory

2. Claims: 56-73 90-96

Mapping of address spaces



## INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 99/30437

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 5123094	A	16-06-1992	NONE	
US 5123101	A	16-06-1992	US 5230045 A	20-07-1993
US 5652885	A	29-07-1997	NONE	
US 5276896	A	04-01-1994	EP 0533787 A WO 9120040 A	31-03-1993 26-12-1991
US 5142683	A	25-08-1992	US 4866664 A AT 78611 T DE 3873054 D DE 3873054 T EP 0321544 A JP 1501660 T WO 8810469 A	12-09-1989 15-08-1992 27-08-1992 11-03-1993 28-06-1989 08-06-1989 29-12-1988
US 5689713	A	18-11-1997	US 5907485 A EP 0738978 A JP 9016542 A US 5892957 A EP 0735476 A JP 9069855 A SG 40839 A	25-05-1999 23-10-1996 17-01-1997 06-04-1999 02-10-1996 11-03-1997 14-06-1997
EP 0444376	A	04-09-1991	DE 69029084 D JP 2587141 B JP 4217059 A US 5357612 A	12-12-1996 05-03-1997 07-08-1992 18-10-1994
US 5233701	A	03-08-1993	JP 1246656 A	02-10-1989
US 5590301	A	31-12-1996	CA 2211083 A EP 0855057 A JP 11512857 T WO 9713191 A	10-04-1997 29-07-1998 02-11-1999 10-04-1997
US 5117350	A	26-05-1992	WO 9007154 A	28-06-1990
US 5687342	A	11-11-1997	NONE	